CASIA-SURF CeFA: A Benchmark for Multi-modal Cross-ethnicity Face Anti-spoofing

Ajian Liu[1*], Zichang Tan[2*], Jun Wan[2†], Sergio Escalera[3], Guodong Guo[4], Stan Z. Li[1,5]
[1]M.U.S.T, Macau; [2]NLPR, CASIA, UCAS, China; [3]CVC, UB, Spain
[4]Baidu Research, China; [5]The Westlake University, China
ajianliu92@gmail.com, {zichang.tan,jun.wan}@nlpr.ia.ac.cn
sergio@maia.ub.es, guoguodong01@baidu.com, Stan.ZQ.Li@westlake.edu.cn

Abstract

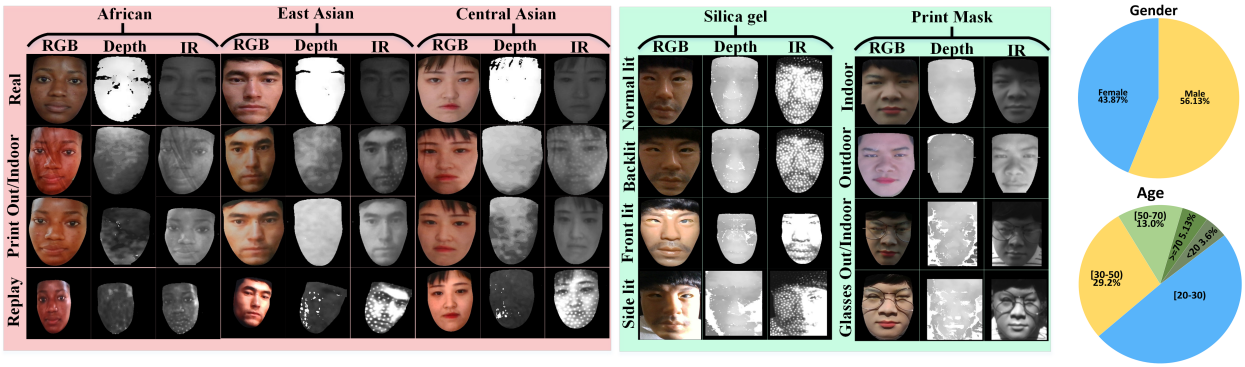
The issue of ethnic bias has proven to affect the performance of face recognition in previous works, while it still remains to be vacant in face anti-spoofing. Therefore, in order to study the ethnic bias for face anti-spoofing, we introduce the largest CASIA-SURF Cross-ethnicity Face Anti-spoofing (CeFA) dataset, covering 3 *ethnicities,* 3 *modalities,* $1,607$ *subjects, and 2D plus 3D attack types. Five protocols are introduced to measure the affect under varied evaluation conditions, such as cross-ethnicity, unknown spoofs or both of them. As our knowledge, CASIA-SURF CeFA is the first dataset including explicit ethnic labels in current released datasets. Then, we propose a novel multi-modal fusion method as a strong baseline to alleviate the ethnic bias, which employs a partially shared fusion strategy to learn complementary information from multiple modalities. Extensive experiments have been conducted on the proposed dataset to verify its significance and generalization capability for other existing datasets,* i.e.*, CASIA-SURF, OULU-NPU and SiW datasets. The dataset is available at* `https://sites.google.com/qq.com/face-anti-spoofing/welcome/challengecvpr2020?authuser=0`.

1. Introduction

Face anti-spoofing (FAS) [5, 19, 22] is a key role to avoid security breaches in face recognition systems. The presentation attack detection (PAD) technique is a vital stage prior to visual face recognition. Although ethnic bias has been verified to severely affect the performance of face recognition systems [1, 4, 24], it still remains to be vacant in face anti-spoofing. Based on the experiment in Section 5.3, the state-of-the-art (SOTA) algorithms also suffer from ethnic bias. More specifically, the value of Attack Presentation Classification Error Rate (ACER) [2] is at least 8% higher in Central Asia than that of East Asia in Table 5. However, there is no available dataset with exactly ethnic labels and protocol for evaluating this bias issue. Furthermore, as shown in Table 1, the existing face anti-spoofing datasets (*i.e*. CASIA-FASD [32], Replay-Attack [7], OULU-NPU [6] and SiW [19]) has limited number of samples and most of them just contain the RGB modality. Although CASIA-SURF [31] is a large dataset in comparison to the existing alternatives, it still provides limited attack types (only 2D print attack) and single ethnicity (East Asia). Therefore, in order to alleviate the above problems, we release the CASIA-SURF CeFA dataset (briefly named CeFA), which is the largest face anti-spoofing dataset up to date in terms of ethnicities, modalities, number of subjects and attack types. The comparisons of current datasets are listed in Table 1. Concretely, attack types of the CeFA dataset are diverse, including printing from cloth, video replay attack, 3D print and silica gel attacks. More importantly, it is the first public dataset designed for exploring the impact of cross-ethnicity. Some original frame of the data sample and the processed sample, *i.e*., keep only face region, are shown in Fig. 1(a).

Moreover, to relieve the ethnic bias, a multi-modal fusion strategy is introduced in this work based on this consideration that indistinguishable real or fake face which is cased by ethnic factors may exhibit quite different properties under other modality. Some fusion methods [31, 20] are published, which restrict the interactions among different modalities since they are independent before the fusion point. Therefore, it is difficult to effectively utilize the modality relatedness from the beginning of the network to its end. In this paper, we propose a Partially Shared Multi-modal Network (PSMM-Net) as a strong baseline to alleviate ethnic and attack pattern bias. On the one hand, it fuses multi-modal features from each feature scale instead of s-

*These authors contributed equally to this work
†Corresponding author



(a) (b)

Figure 1: (a): Processed samples of the CeFA dataset. It contains $1,607$ subjects, 3 different ethnicities (*i.e.*, Africa, East Asia, and Central Asia), with 4 attack types (*i.e.*, print attack, replay attack, 3D print and silica gel attacks) under various lighting conditions. Light red/blue background indicates 2D/3D attack. (b): Gender and age distributions of the CeFA.

Table 1: Comparisons among existing face PAD databases. (*i* indicates the dataset only contains images. * indicates the dataset contains 4 ethnicities, while it does not provide accurate ethnic labels for each sample and does not study ethnic bias for the design protocol. AS: Asian, A: Africa, U: Caucasian, I: Indian, E: East Asia, C: Central Asia.)

Dataset	Year	#Subject	#Num	Attack	Modality	Device	Ethnicity
Replay-Attack [7]	2012	50	1200	Print,Replay	RGB	RGB Camera	-
CASIA-FASD [32]	2012	50	600	Print,Cut,Replay	RGB	RGB Camera	-
3DMAD [10]	2014	17	255	3D print mask	RGB/Depth	RGB Camera/Kinect	-
MSU-MFSD [26]	2015	35	440	Print,Replay	RGB	Cellphone/Laptop	-
Replay-Mobile [9]	2016	40	1030	Print,Replay	RGB	Cellphone	-
Msspoof [8]	2016	21	4704^i	Print	RGB/IR	RGB/IR Camera	-
OULU-NPU [6]	2017	55	5940	Print,Replay	RGB	RGB Camera	-
SiW [19]	2018	165	4620	Print,Replay	RGB	RGB Camera	AS/A/ U/I*
CASIA-SURF [31]	2019	1000	21000	Print,Cut	RGB/Depth/IR	Intel Realsense	E
CeFA (Ours)	2019	1500	18000	Print, Replay	RGB/Depth/IR	Intel Realsense	A/E/C
		99	5346	3D print mask			
		8	192	3D silica gel mask			
		Total: 1607 subjects, 23538 videos					

tarting from a certain fusion point [31, 20]. On the other hand, it allows the information exchanges and interactions among different modalities by introducing a shared branch. In addition, for each single-modal branch (*e.g.*, RGB, Depth or IR), we use a simple and effective backbone, Resnet [15], to learn the static features for subsequent feature fusion.

To sum up, the contributions of this paper are summarized as follows: (1) We release the largest face anti-spoofing dataset CeFA up to date, which includes 3 ethnicities, 1607 subjects and 4 diverse 2D/3D attack types. (2) We provide a benchmark with five comprehensive evaluation protocols to measure ethnic and attack pattern bias. (3) We propose the PSMM-Net as a strong baseline to alleviate the ethnic bias. (4) Extensive experiments have been conducted on the proposed dataset to verify its significance.

2. Related work

2.1. Datasets

Several studies [13, 17, 21, 24] have uncovered ethnicity bias in face recognition algorithms, and Wang *et al*. [24] has collected a face recognition dataset containing 4 ethnicities used for algorithm design to eliminate ethnicity bias. However, there is no publicly available face anti-spoofing dataset with ethnic labels for research this issue in face anti-spoofing. One can see the following deficiencies from Table 1 which lists existing face anti-spoofing datasets: (1) The maximum number of available subjects was 165 on the SiW dataset [19] before 2019; (2) Most of the datasets just contain RGB data, such as Replay-Attack [7], CASIA-FASD [32], SiW [19] and OULU-NPU [6]; (3)

Most datasets do not provide ethnicity information, except SiW and CASIA-SURF. Although SiW provides four ethnicities, it has neither a clear ethnic label nor a standard protocol for measuring ethnic bias in algorithms. This limitation also holds for the CASIA-SURF dataset.

2.2. Methods

VIS-based Methods. Since most FAS systems adopt RGB camera, a considerable part of face PAD methods [19, 27, 25, 28] were designed in VIS spectrum. Therefore, the color texture information is an important clues for FAS task. Recently, some works [11, 18] attempts to learn CNN-based features by utilizing deep learning framework in an end-to-end manner. Concurrent to the supervision of using softmax loss, another works derive inspiration from physical cues, that establish a commonality for genuine face and distinction from fake ones. Liu *et al*. [19] design a CNN-RNN model to leverage Depth map and rPPG signal as supervision. In this work, we employ a simple and effective Resnet [15] as baseline to learn the static texture feature.

Multi-modal Fusion Methods. Zhang *et al*. [31] proposed a fusion network with 3 streams using Resnet-18 as the backbone, where each stream is used to extract low level features from RGB, Depth and IR data, respectively. All previous methods just consider as a key fusion component the concatenation of features from multiple modalities. Unlike [31, 20, 23], we propose the PSMM-Net, where three modality-specific networks and one shared network are connected by using a partially shared structure to learn discriminative fused features for face anti-spoofing.

3. CeFA dataset

In this section, we introduce the CeFA dataset in detail, such as acquisition details, attack types, and protocols.

Acquisition Details. We use the Intel Realsense to capture the RGB, Depth and IR videos simultaneously at $30fps$. The resolution is 1280×720 pixels for each frame in video. Subjects are asked to move smoothly their head so as to have a maximum of around 30^0 deviation of head pose in relation to frontal view. Data pre-processing is similar to the one performed in [31], expect that PRNet [12] is replaced by 3DDFA [33, 14] for face region detection.

Statistics. As shown in Table 1, CeFA consists of 2D and 3D attack subsets. As shown in Fig. 1(a), for the 2D attack subset, it consists of print and video-replay attacks captured by subjects from three ethnicites (*e.g*., African, East Asian and Central Asian). See from the Table 2, each ethnicity has 500 subjects, and each subject has 1 real sample, 2 fake samples of print attack captured in indoor and outdoor, and 1 fake sample of video-replay. In total, there are $18,000$ videos ($6,000$ per ethnicity). The age and gender statistics for the 2D attack subset of CeFA is shown in Fig. 1(b).

Table 2: Statistics of the 2D attack subset of the CeFA.

Ethnicity	Real & Attack styles	# RGB	# Depth	# IR	Subtotal
African East Asian Central Asian	Real	500	500	500	6000 6000 6000
	Cloth-indoor attack	500	500	500	
	Cloth-outdoor attack	500	500	500	
	Replay attack	500	500	500	
Total: 1500 subjects, 18000 videos					

Table 3: Statistics of the 3D attack subset of the CeFA.

3D Mask Attack	Attack styles	# RGB	# Depth	# IR	Subtotal
Print mask 99 Subjects & 6 Lighting	Only mask	594	594	594	5346
	Wig without glasses	594	594	594	
	Wig with glasses	594	594	594	
Silica gel mask 8 Subjects & 4 Lighting	Wig without glasses	32	32	32	192
	Wig with glasses	32	32	32	
Total: 107 subjects, 5538 videos					

For the 3D attack subset in Table 3, it has 3D print mask and silica gel face attacks. Some samples are shown in Fig. 1(a). In the part of 3D print mask, it has 99 subjects, each subject with 18 fake samples captured in three attacks and six lighting environments. Specially, attack types include only face mask, wearing a wig with glasses, and wearing a wig without glasses. Lighting conditions include outdoor sunshine, outdoor shade, indoor side light, indoor front light, indoor backlit and indoor regular light. In total, there are $5,346$ videos ($1,782$ per modality). For silica gel face attacks, it has 8 subjects, each subject has 8 fake samples captured in two attacks styles and four lighting environments. Attacks include wearing a wig with glasses and wearing a wig without glasses. Lighting environments include indoor side light, indoor front light, indoor backlit and indoor normal light. In total, there are 192 videos (64 per modality).

Evaluation Protocols. The motivation of CeFA dataset is to provide a benchmark to measure the generalization performance of new PAD methods in three main aspects: cross-ethnicity, cross-modality, cross-attacks, and the fairness of PAD methods in different ethnicities. We design five protocols for the 2D attacks subset, as shown in Table 4, totalling 12 sub-protocols (1_1, 1_2, 1_3, 2_1, 2_2, 3_1, 3_2, 3_3, 4_1, 4_2, 4_3, and 5). We divide 500 subjects per ethnicity into three subject-disjoint subsets (second and fourth columns in Table 4). Each protocol has three data subsets: training, validation and testing sets, which contain 200, 100, and 200 subjects, respectively.

• **Protocol 1 (cross-ethnicity)**: Most of the public face PAD datasets lack of ethnicity labels or do not provide with a protocol to perform cross-ethnicity evaluation. Therefore, we design the first protocol to evaluate the generalization of PAD methods for cross-ethnicity testing. One ethnicity is used for training and validation, and the left two ethnicities

Table 4: Five protocols are defined for CeFA: (1) cross-ethnicity, (2) cross-PAI, (3) cross-modality, (4) cross-ethnicity&PAI, (5) bias-ethnicity

. Note that the 3D attacks subset are included in each testing protocol (not shown in the table). & indicates merging; $*\_*$ corresponds to the name of sub-protocols. R: RGB, D: Depth, I: IR. Other abbreviated as in Table 1.

Prot.	Subset	Ethnicity			Subjects	Modalities			PAIs		# real videos	# fake videos	# all videos
		1_1	1_2	1_3									
1	Train	A	C	E	1-200	R&D&I			Print&Replay		600/600/600	1800/1800/1800	2400/2400/2400
	Valid	A	C	E	201-300	R&D&I			Print&Replay		300/300/300	900/900/900	1200/1200/1200
	Test	C&E	A&E	A&C	301-500	R&D&I			Print&Replay		1200/1200/1200	6600/6600/6600	7800/7800/7800
									2_1	2_2			
2	Train	A&C&E			1-200	R&D&I			Print	Replay	1800/1800	3600/1800	5400/3600
	Valid	A&C&E			201-300	R&D&I			Print	Replay	900/900	1800/900	2700/1800
	Test	A&C&E			301-500	R&D&I			Replay	Print	1800/1800	4800/6600	6600/8400
						3_1	3_2	3_3					
3	Train	A&C&E			1-200	R	D	I	Print&Replay		600/600/600	1800/1800/1800	2400/2400/2400
	Valid	A&C&E			201-300	R	D	I	Print&Replay		300/300/300	900/900/900	1200/1200/1200
	Test	A&C&E			301-500	D&I	R&I	R&D	Print&Replay		1200/1200/1200	5600/5600/5600	6800/6800/6800
		4_1	4_2	4_3									
4	Train	A	C	E	1-200	R&D&I			Replay		600/600/600	600/600/600	1200/1200/1200
	Valid	A	C	E	201-300	R&D&I			Replay		300/300/300	300/300/300	600/600/600
	Test	C&E	A&E	A&C	301-500	R&D&I			Print		1200/1200/1200	5400/5400/5400	6600/6600/6600
		5											
5	Train	A&C&E			1-200	R&D&I			Print&Replay		1800	5400	7200
	Valid	A&C&E			201-300	R&D&I			Print&Replay		900	2700	3600
	Test	A	C	E	301-500	R&D&I			Print&Replay		600/600/600	3800/3800/3800	4400/4400/4400

are used for testing. Therefore, there are three different evaluations (third column of Protocol 1 in Table 4).

• **Protocol 2 (cross-PAI)**: Given the diversity and unpredictability of attack types from different presentation attack instruments (PAI), it is necessary to evaluate the robustness of face PAD algorithms to this kind of variations (sixth column of Protocol 2 in Table 4).

• **Protocol 3 (cross-modality)**: Inspired by heterogeneous face recognition, we define three cross-modality evaluations, each of them having one modality for training and the two remaining ones for testing (fifth column of Protocol 3 in Table 4). Although there are no real world scenarios for this protocol until now, if algorithms trained on a certain modality data are able to perform well on other modalities data, this will greatly enhance their versatility for different scenes with different devices. Similar to [30], we aim to provide this cross-modal evaluation protocol for those possible real-world scenarios in the future.

• **Protocol 4 (cross-ethnicity & PAI)**: The most challenging protocol is designed via combining the condition of both Protocol 1 and 2. As shown in Protocol 4 of Table. 4, the testing subset introduces two unknown target variations simultaneously.

• **Protocol 5 (bias-ethnicity)**: Algorithm fairness has started to attract the attention of researchers in Artificial Intelligence (AI). According to this criterion: an ideally fair algorithm should have consistent performance on different protected attributes. In this paper, in addition to measuring the generalization performance of the new methods on cross-ethnicity (*i.e.*, Protocol 1), we also consider the fairness of an algorithm, where it is trained with data that includes all ethnicities, and assessed on different ethnicities, respectively. Like [6], the mean and variance of evaluate metrics for five protocols are calculated in our experiments. Detailed statistics for the different protocols are shown in Table 4.

4. Proposed Method

First, a simple and effective Resnet [15] is employed in this work to learn the static texture features for each modality. It consists of 5 blocks (*i.e.*, conv, res1, res2, res3, res4) and 1 Global Average Pooling (GAP) layer. Then, the PSMM-Net is presented by learning the fusion features from multiple modalities.

4.1. PSMM-Net for Multi-modal Fusion

The architecture of the proposed PSMM-Net is shown in Fig. 2. It consists of two main parts: a) the modality-specific network, which contains three Resnet-18 [15] to learn features from RGB, Depth, IR modalities, respectively; b) and a shared branch for all modalities, which aims to learn the complementary features among different modalities. For the shared branch, we adopt Resnet-18, removing the first conv layer and res1 block. In order to capture correlations and complementary semantics among different modalities, information exchange and interaction among modality-specific branches and the shared branch are designed. This is done in two different ways: a) forward feeding (*i.e.*, black arrow) of fused modality-specific features to the shared branch, and b) backward feeding (*i.e.*, light green arrow) from shared branch modules output to modality-specific block inputs.

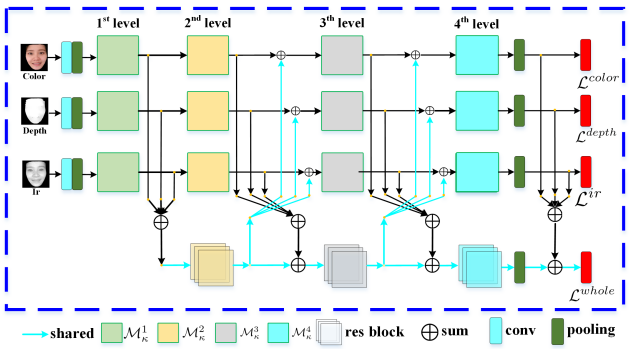


Figure 2: PSMM-Net diagram consists of two main parts: (1) Modality-specific network, which contains three Resnets; (2) A shared branch for all modalities, which aims to learn the complementary features among different modalities. we divide residual blocks of the modality-specific branch into a set of modules $\{\mathcal{M}_\kappa^t\}_{t=1}^4$ according to feature level, where $\kappa \in \{color, depth, ir\}$ is an indicator of the modality and t represents the feature level.

Forward Feeding. We fuse modality-specific features from all modality branches and feed them as input to its corresponding shared block. The fused process at t^{th} feature level can be formulated as:

$$\tilde{\mathbf{S}}^t = \sum_\kappa \mathbf{X}_\kappa^t + \mathbf{S}^t \quad t = 1, 2, 3 \tag{1}$$

where $\mathbf{X}_k^t$ is the output of the modality-specific block, $k \in \{color, depth, ir\}$ is an indicator of the modality and t represents the feature level. In the shared branch, $\tilde{\mathbf{S}}^t$ denotes the input to the $(t+1)^{th}$ block, and $\mathbf{S}^t$ denotes the output of the t^{th} block. Note that the first residual block is removed from the shared branch, thus $\mathbf{S}^1$ equals to zero.

Backward Feeding. Shared features $\mathbf{S}^t$ are delivered back to the modality-specific networks. The static texture features $\mathbf{X}_\kappa^t$ add with $\mathbf{S}^t$ for feature fusion. This can be denoted as:

$$\widetilde{\mathbf{X}}_\kappa^t = \mathbf{X}_\kappa^t + \mathbf{S}^t, \quad t = 2, 3 \tag{2}$$

After feature fusion, $\widetilde{\mathbf{X}}_\kappa^t$ become the new features, which are then feed to the next module $\mathcal{M}_\kappa^{t+1}$.

Loss Optimization. The binary cross-entropy loss is used as the loss function. In summary, there are two kinds of losses employed to guide the training of PSMM-Net. The first corresponds to the losses of the three modality-specific branches, *i.e.* color, depth and ir modalities, denoted as $\mathcal{L}^{color}$, $\mathcal{L}^{depth}$ and $\mathcal{L}^{ir}$, respectively. The second corresponds to the loss that guides the entire network training, denoted as $\mathcal{L}^{whole}$, which bases on the summed features from all branches. The overall loss $\mathcal{L}$ of PSMM-Net is denoted as:

$$\mathcal{L} = \mathcal{L}^{whole} + \mathcal{L}^{color} + \mathcal{L}^{depth} + \mathcal{L}^{ir} \tag{3}$$

5. Experiments

In this section, we conduct a series of experiments on CeFA and public available face anti-spoofing datasets to show the significance of the presented dataset and generalization capability.

5.1. Datasets & Metrics

We evaluate the performance of PSMM-Net on two multi-modal (*i.e.*, RGB, Depth and IR) datasets: CeFA and CASIA-SURF [31], while evaluate the modality-specific network on two single-modal (*i.e.*, RGB) face anti-spoofing benchmarks: OULU-NPU [6] and SiW [19]. Similar to [31], experiments on other datasets only verify the generalization performance of the proposed CeFA by setting the with/without of CeFA as pre-training. In order to perform a consistent evaluation with prior works, we report the experimental results using the following metrics based on respective official protocols: Attack Presentation Classification Error Rate (APCER) [2], Bona Fide Presentation Classification Error Rate (BPCER), Average Classification Error Rate (ACER), and Receiver Operating Characteristic (ROC) curve [31].

Inspired by the competition of "Looking at People Fair Face Recognition challenge ECCV2020[1]", the participants will be asked to develop their fair face verification method aiming for a reduced bias in terms of gender and skin color (protected attributes). Before illustrating the definitions of fairness in this work, we checked whether the method that uses this dataset exhibits ethnicity-related bias via calculating the value of $Bias_{EER}$:

$$Bias_{EER} = \sum_e ERR_e - min_{e'} ERR_{e'} \tag{4}$$

where ERR denotes the error metric, such as APCER, BPCER, or ACER, e represents the ethnicity in CeFA, such as AF, CA, or EA, $e^{'}$ is the ethnicity with the lowest ERR, and $Bias_{EER}$ means the total bias (non-negative value) of the algorithm in one metric. Informally, we define an algorithm as fair if it achieves the same error for all protected ethnicities under the metric of ACER.

5.2. Implementation Details

The proposed PSMM-Net is implemented with Tensorflow [3] and run on a single NVIDIA TITAN X GPU. We resize the cropped face region to 112×112, and use random

[1] http://chalearnlap.cvc.uab.es/challenge/38/description/

rotation within the range of $[-30^0, 30^0]$, flipping, cropping and color distortion for data augmentation. All models are trained for 25 epochs via Adaptive Moment Estimation (Adam) algorithm and initial learning rate of 0.1, which is decreased after 15 and 20 epochs with a factor of 10. The batch size of each CNN stream is 64.

5.3. Performance Biases of Diversity Ethnicities

In this section, we investigate the performance biases of different ethnicities when two SOTA algorithms on the three ethnicities of our CeFA, respectively. The MS-SEF [31] is trained on CASIA-SURF for the multi-modal data while FAS-BAS [19] is trained for the RGB data on OULU-NPU. Then, the trained models are tested on CeFA. The results are shown in Table 5. It shows that the results of both methods is different among three ethnicities, such as East Asian (11.4%) versus Center Asian (19.6%) for MS-SEF and African (14.2%) versus Center Asian (26.1%) for MS-SEF under the ACER metric. In addition, we found the two methods that achieved relatively good results on East Asians (*e.g.*, the values of ACER are 11.4%, 15.4%, respectively) due to the most of samples belong to East Asians on CASIA-SURF and OULU-NPU. It indicates that the existing single-ethnic anti-spoofing datasets limit the ethnic generalization performance of existing methods.

5.4. Baseline Model Evaluation

Before exploring the traits of our dataset, we first provide a benchmark for CeFA based on the proposed method. From the Table 6, we can draw the following conclusions: (1) The ACER scores of three sub-protocols in Protocol 1 are 2.3%, 4.8% and 3.4%, respectively, which indicating the necessity to study the generalization of the face PAD methods for different ethnicities; (2) In the case of Protocol 2, when print attack is used for training/validation and video-replay and 3D mask are used for testing, the ACER score is 1.6% (sub-protocol 2_1), while video-replay attack is used for training/validation, and print attack and 3D attack are used for testing, with an ACER score of 9.1% (sub-protocol 2_2). The large gap between the results caused by the different PAI (*i.e.*, different displays and printers); (3) Protocol 3 evaluates cross-modality. The best result is achieved for sub-protocol 3_1 (ACER=6.2%); (4) Protocol 4 is the most difficult evaluation scenario, which simultaneously considers cross-ethnicity and cross-PAI. All sub-protocols achieve poor performance which highlighting the challenge of our dataset, being 4.2%, 8.4%, and 7.6% ACER scores for sub-protocols of 4_1, 4_2, and 4_3, respectively; (5) In order to measure the fairness of the algorithm, we first train a model with a training set that combines three ethnicities (*i.e.*, AF, CE, EA), then evaluate its performance on different ethnicities based on the model, and finally calculate the bias of the model according to formula 4. It can be seen from Protocol 5 in Table 6 that our baseline method has better performance in terms of ACER compared to other protocols because the training set contains data for all ethnicities. However, discrimination still exists on the three ethnicities, with biases of 0.6, 0.3, 0.5 for APCER, BPCER, and ACER, respectively.

Table 6: PSMM-Net evaluation on the five protocols of CeFA dataset, where A_B represents sub-protocol B from Protocol A, and Avg±Std indicates the mean and variance operation.

Prot. name		APCER(%)	BPCER(%)	ACER(%)
Prot. 1	1_1	1.7	2.8	2.3
	1_2	2.5	7.1	4.8
	1_3	2.9	3.8	3.4
	Avg±Std	2.4±0.6	4.6±2.3	3.5±1.3
Prot. 2	2_1	1.3	1.9	1.6
	2_2	14.0	4.2	9.1
	Avg±Std	7.7±9.0	3.1±1.6	5.4±5.3
Prot. 3	3_1	9.5	2.9	6.2
	3_2	24.3	6.2	15.3
	3_3	24.5	5.9	15.2
	Avg±Std	19.4±8.7	5.0±1.8	12.2±5.2
Prot. 4	4_1	5.0	3.3	4.2
	4_2	7.7	9.0	8.4
	4_3	10.8	4.3	7.6
	Avg±Std	7.8±2.9	5.5±3.0	6.7±2.2
Prot. 5	AF	1.2	1.4	1.3
	CA	1.4	1.5	1.5
	EA	1.6	1.6	1.6
	Bias	0.6	0.3	0.5

5.5. Ablation Analysis

To verify the performance of our proposed baseline in alleviating ethnic bias, we perform a series of ablation experiments on Protocol 1 (cross-ethnicity) of the CeFA dataset.

Multiple Modalities. In order to show the effect of analysing a different number of modalities, we evaluate one modality (RGB), two modalities (RGB and Depth), and three modalities (RGB, Depth and IR) on PSMM-Net. As shown in Fig. 2, the PSMM-Net contains three modality-specific branches and one shared branch. When only RGB modality is considered, we just use one Resnet for evaluation. When two or three modalities are considered, we use two or three Resnets and one shared branch to train the PSMM-Net model, respectively. Results are shown in Table 7. The best results are obtained when using all three modalities, which 2.4% of APCER, 4.6% of BPCER and 3.5% of ACER. The comparison results show that the multi-modal information has a significant effect in alleviating the issue of ethnic bias, which is mainly due to the smaller differences in skin color of different ethnicities in the IR modality.

Table 5: Ethnic bias in SOTA PAD methods. The ACER(%) on three ethnicities of proposed CeFA are given.

Method	Trained Dataset	Modality	Ethnicity(ACER%)		
			Africa	Central Asia	East Asia
MS-SEF [31]	CASIA-SURF [31]	RGB&Depth&IR	13.9	19.6	11.4
FAS-BAS [19]	OULU-NPU [6]	RGB	14.2	26.1	15.4

Table 7: Effect of multiple modalities.

Prot.1	PSMM-Net		
	APCER(%)	BPCER(%)	ACER(%)
RGB	15.7±5.3	12.4±2.2	14.1±3.8
RGB&Depth	5.2±2.3	13.6±5.2	9.4±3.1
RGB&Depth&IR	**2.4±0.6**	**4.6±2.3**	**3.5±1.3**

Table 8: Comparison of fusion strategies.

Method	APCER(%)	BPCER(%)	ACER(%)
NHF	25.3±12.2	4.4±3.1	14.8±6.8
PSMM-WoBF	12.7±0.4	3.2±2.3	7.9±1.3
PSMM-Net	**2.4±0.6**	**4.6±2.3**	**3.5±1.3**

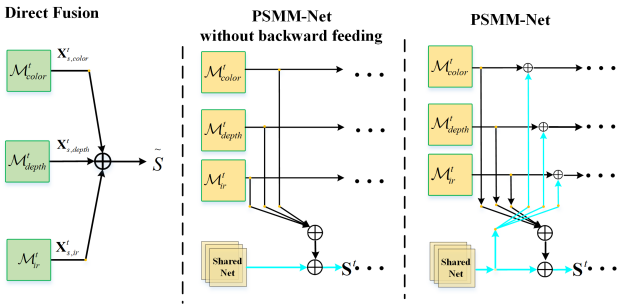


Figure 3: Comparison of network units for multi-modal fusion strategies. From left to right: NHF, PSMM-NET-WoBF and PSMM-Net. The fusion process for the t^{th} feature level of each strategy is shown at the bottom.

Fusion Strategy. In order to evaluate the performance of PSMM-Net, we compare it with other two variants: Naive halfway fusion (NHF) and PSMM-Net without backward feeding mechanism (PSMM-Net-WoBF). As shown in Fig. 3, NHF combines the modules of different modalities at a later stage (*i.e.*, after $\mathcal{M}_{\kappa}^{1}$ module) and PSMM-Net-WoBF strategy removes the backward feeding from PSMM-Net. The fusion comparison results are shown in Table 8, showing higher performance of the proposed PSMM-Net with information exchange and interaction mechanism among modality-specific branches and the shared branch.

5.6. Using CeFA for Pre-Training

In this section, the PSMM-Net and Resnet are adopted as the baseline to evaluate the generalization of the proposed dataset on multi-modal dataset, *i.e.*, CASIA-SURF and single-modal datasets, *i.e.*, OULU-NPU and SiW, respectively. Similar to [30], we first pre-train the model on CeFA and then fine-tune with the concerned datasets, which is termed as PSMM-Net (CeFA) or Resnet (CeFA).

CASIA-SURF. It is a large publicly available dataset for face anti-spoofing in terms of both subjects and modalities. Based on the official protocol [31], we compare with three methods to demonstrate the superiority of our PSMM-Net and the generalization capability of proposed CeFA dataset. From the results which are show in Table 9, we can see the performance of the PSMM-Net is superior to the ones of the competing multi-modal fusion methods, including Halfway fusion [31], single-scale SE fusion [31], and multi-scale SE fusion [29]. When compared with [31, 29], PSMM-Net improves the performance by at least 0.4% for ACER. When the PSMM-Net is pretrained on CeFA, it further improves the performance. Concretely, the performance of $TPR@FPR = 10^{-4}$ is increased by 2.4% when pretraining with the proposed CeFA dataset. The comparison results not only illustrate the superiority of our algorithm for multi-modal data fusion, but also show that our CeFA alleviates the bias of attack pattern to a certain extent.

OULU-NPU. See from the Table 1, it is a high-resolution dataset, consisting of $5,940$ videos corresponding to 55 subjects recorded in three different illumination conditions. There are 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. We compare the Resnet with other SOTA methods, *i.e.*, BAS [19], Ds [16], STASN [27]. From the results in Table 10, our method which is pre-trained by proposed dataset achieves the best results (The lower ACER value indicates the better performance) on protocol 2, 3 and 4 of the OULU-NPU. Especially in the most difficult Protocol 4, using the proposed dataset to pre-train our baseline method significantly improves its ACER performance, *i.e.*, from 12.0% to

Table 9: Comparison of the proposed method with three fusion strategies. All models are trained and tested on the CASIA-SURF. '()' means the method is pre-trained with a specific dataset. Best results are bolded.

Method	TPR (%)			APCER (%)	BPCER (%)	ACER (%)
	@FPR=10^{-2}	@FPR=10^{-3}	@FPR=10^{-4}			
NHF [31]	89.1	33.6	17.8	5.6	3.8	4.7
Single-scale SEF [31]	96.7	81.8	56.8	3.8	1.0	2.4
Multi-scale SEF [29]	99.8	98.4	95.2	1.6	0.08	0.8
PSMM-Net	**99.9**	99.3	96.2	0.7	0.06	0.4
PSMM-Net(CeFA)	**99.9**	**99.7**	**97.6**	**0.5**	**0.02**	**0.2**

Table 10: Comparisons on OULU-NPU.

Pro.	Method	APCER(%)	BPCER(%)	ACER(%)
1	BAS [19]	1.6	1.6	1.6
	Ds [16]	1.2	1.7	1.5
	STASN [27]	1.2	2.5	1.9
	Resnet	**0.8**	**4.2**	**2.5**
	Resnet(CeFA)	**1.7**	**1.7**	**1.7**
2	BAS	2.7	2.7	2.7
	STASN	4.2	0.3	2.2
	Resnet	**4.0**	**1.9**	**3.0**
	Resnet(CeFA)	**1.4**	**2.5**	**2.0**
3	BAS	2.7±1.3	3.1±1.7	2.9±1.5
	STASN	4.7±3.9	0.9±1.2	2.8±1.6
	Resnet	**3.5±2.4**	**4.7±2.1**	**4.1±2.3**
	Resnet(CeFA)	**2.3±1.5**	**3.2±1.7**	**2.8±1.4**
4	BAS	9.3±5.6	10.4±6.0	9.5±6.0
	STASN	6.7±10.6	8.3±8.4	7.5±4.7
	Resnet	**12.3±4.7**	**11.7±5.2**	**12.0±5.5**
	Resnet(CeFA)	**6.4±3.6**	**7.2±4.1**	**6.8±4.3**

6.8%. It reveals that our CeFA can alleviate the bias issue of the acquisition device and attack type of the PAD algorithm to a certain extent.

SiW. It provides live and spoof videos from 165 subjects. In addition, they provide three protocols for future study on SiW. Table 11 shows the comparison between our method with three SOTA methods, *i.e.*, BAS [19], TD-SF [25] and STASN [27]. Similar conclusions in the OULU-NPU experiment, our pre-trained Resnet on CeFA can achieve the best results on all protocols. Compared with the method of Resnet, the performance of ACER is reduced by 3.06%, 0.59% and 2.75% in Protocol 1, 2, and 3 respectively when using the proposed CeFA dataset as pre-training.

In summary, we believe that other SOTA methods can be further improved by using our CeFA as the pre-training dataset. Those experimental results clearly demonstrate the effectiveness and generalization capability of the collected CeFA dataset.

Table 11: Comparisons on SiW. 'Pro.' denotes the protocol.

Pro.	Method	APCER(%)	BPCER(%)	ACER(%)
1	BAS [19]	3.58	3.58	3.58
	TD-SF [25]	1.27	0.83	1.05
	STASN [27]	-	-	1.00
	Resnet	**1.79**	**6.18**	**3.99**
	Resnet(CeFA)	**1.03**	**0.83**	**0.93**
2	BAS	0.57±0.69	0.57±0.69	0.57±0.69
	TD-SF	0.33±0.27	0.29±0.39	0.31±0.28
	STASN	-	-	0.28±0.05
	Resnet	**0.75±0.22**	**0.89±0.32**	**0.82±0.23**
	Resnet(CeFA)	**0.20±0.11**	**0.25±0.22**	**0.23±0.15**
3	BAS	8.31±3.81	8.31±3.81	8.31±3.81
	TD-SF	7.70±3.88	7.76±4.09	7.73±3.99
	STASN	-	-	12.10±1.50
	Resnet	**9.46±4.21**	**9.12±4.55**	**9.29±4.27**
	Resnet(CeFA)	**6.35±3.67**	**6.72±3.75**	**6.54±3.46**

6. Conclusion

In this paper, we release the largest face anti-spoofing dataset up to date in terms of modalities, number of subjects and attack types. More importantly, CeFA is the only public face anti-spoofing dataset with ethnic label. In addition, we provide a baseline, namely PSMM-Net, by learning complementary information from multi-modal data to alleviate the ethnic bias. Extensive experiments validate the utility of our algorithm and the generalization capability of models trained on the proposed dataset.

7. Acknowledgments

This work was supported by the Chinese National Natural Science Foundation Projects #61961160704, #61876179, Science and Technology Development Fund of Macau (No. 0025/2018/A1, 0010/2019/AFJ), the Key Project of the General Logistics Department Grant No. ASW17C001, the Spanish project PID2019-105093GB-I00 (MINECO/FEDER, UE) and CERCA Programme / Generalitat de Catalunya, and by ICREA under the ICREA Academia programme. We acknowledge Surfing Technology Beijing co., Ltd (www.surfing.ai) to provide high-quality dataset.

References

[1] Are face recognition systems accurate? depends on your race. 2016. https://www.technologyreview.com/s/601786.

[2] ISO/IEC JTC 1/SC 37 Biometrics. information technology biometric presentation attack detection part 1: Framework. international organization for standardization. 2016. https://www.iso.org/obp/ui/iso.

[3] Martłn Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, and Xiaoqiang Zhang. Tensorflow: A system for large-scale machine learning. 2016.

[4] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellaker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings.

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

[6] 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.

[7] I. Chingovska, A. Anjos, and S. Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *Biometrics Special Interest Group*, 2012.

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

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

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

[11] 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.

[12] 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.

[13] Nicholas Furl, P. Jonathon Phillips, and Alice J. O'Toole. Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis. In *Cognitive science*, 2002.

[14] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *ECCV*, 2020.

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

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

[17] Brendan F. Klare, Mark J. Burge, Joshua C. Klontz, Richard W. Vorder Bruegge, and Anil K. Jain. Face recognition performance: Role of demographic information. volume 7, pages 1789–1801.

[18] 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.

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

[20] Aleksandr Parkin and Oleg Grinchuk. Recognizing multi-modal face spoofing with face recognition networks. In *PRCVW*, pages 0–0, 2019.

[21] P. Jonathon Phillips, Jiang Fang, Abhijit Narvekar, Julianne H. Ayyad, and Alice J. O'Toole. An other-race effect for face recognition algorithms. volume 8, page 14, 2011.

[22] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *CVPR*, pages 10023–10031, 2019.

[23] Tao Shen, Yuyu Huang, and Zhijun Tong. Facebagnet: Bag-of-local-features model for multi-modal face anti-spoofing. In *PRCVW*, pages 0–0, 2019.

[24] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *ICCV*, October 2019.

[25] Zezheng Wang, Zitong Yu, Chenxu Zhao, Xiangyu Zhu, Yunxiao Qin, Qiusheng Zhou, Feng Zhou, and Zhen Lei. Deep spatial gradient and temporal depth learning for face anti-spoofing. *CVPR*, 2020.

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

[27] Xiao Yang, Wenhan Luo, Linchao Bao, Yuan Gao, Dihong Gong, Shibao Zheng, Zhifeng Li, and Wei Liu. Face anti-spoofing: Model matters, so does data. In *CVPR*, pages 3507–3516, 2019.

[28] Zitong Yu, Chenxu Zhao, Zezheng Wang, Yunxiao Qin, Zhuo Su, Xiaobai Li, Feng Zhou, and Guoying Zhao. Searching central difference convolutional networks for face anti-spoofing. *CVPR*, 2020.

[29] Shifeng Zhang, Ajian Liu, Jun Wan, Yanyan Liang, Guogong Guo, Sergio Escalera, Hugo Jair Escalante, and Stan Z Li. Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing. *arXiv:1908.10654*, 2019.

[30] Shifeng Zhang, Ajian Liu, Jun Wan, Yanyan Liang, and Stan Z. Li. Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing. *TBMIO*, 2019.

[31] Shifeng Zhang, Xiaobo Wang, Ajian Liu, Chenxu Zhao, Jun Wan, Sergio Escalera, Hailin Shi, Zezheng Wang, and Stan Z. Li. A dataset and benchmark for large-scale multi-modal face anti-spoofing. In *CVPR*, 2019.

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

[33] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z Li. Face alignment in full pose range: A 3d total solution. *TPAMI*, 41(1):78–92, 2017.