Learning Polysemantic Spoof Trace: A Multi-Modal Disentanglement Network for Face Anti-spoofing

Kaicheng Li[1, 2], Hongyu Yang[3*], Binghui Chen, Pengyu Li, Biao Wang, Di Huang[1, 2, 4]

[1]State Key Laboratory of Software Development Environment, Beihang University, China
[2]School of Computer Science and Engineering, Beihang University, China
[3]Institute of Artificial Intelligence, Beihang University, China
[4]Hangzhou Innovation Institute, Beihang University, China
{likaicheng, hongyuyang, dhuang}@buaa.edu.cn, chenbinghui@bupt.edu.cn, lipengyu007@gmail.com,
wangbiao225@foxmail.com

Abstract

Along with the widespread use of face recognition systems, their vulnerability has become highlighted. While existing face anti-spoofing methods can be generalized between attack types, generic solutions are still challenging due to the diversity of spoof characteristics. Recently, the spoof trace disentanglement framework has shown great potential for coping with both seen and unseen spoof scenarios, but the performance is largely restricted by the single-modal input. This paper focuses on this issue and presents a multi-modal disentanglement model which targetedly learns polysemantic spoof traces for more accurate and robust generic attack detection. In particular, based on the adversarial learning mechanism, a two-stream disentangling network is designed to estimate spoof patterns from the RGB and depth inputs, respectively. In this case, it captures complementary spoofing clues inhering in different attacks. Furthermore, a fusion module is exploited, which recalibrates both representations at multiple stages to promote the disentanglement in each individual modality. It then performs cross-modality aggregation to deliver a more comprehensive spoof trace representation for prediction. Extensive evaluations are conducted on multiple benchmarks, demonstrating that learning polysemantic spoof traces favorably contributes to anti-spoofing with more perceptible and interpretable results.

Introduction

Face Recognition (FR) has been universally accepted and employed in our daily life, involving a diversity of applications, such as mobile payment and access control. But along with its fast popularization, face Spoofing Attacks (SAs, also known as presentation attacks) have become an increasingly critical issue. Generally, the attackers can grab faces of target individuals with cameras or from social networks, which are then counterfeited as reprinted photos, replayed videos or 3D masks for invasion. FR systems encounter enormous threats when attackers masquerade as others. Correspondingly, Face Anti-Spoofing (FAS) is essential.

In the literature, an overwhelming majority of the studies detect SAs with RGB images or video frames, where they mainly perform analysis on the texture difference between the genuine faces and the spoofing ones, *e.g.,* color distortions, unnatural specular highlights, and Moiré patterns. The hand-crafted features, such as LBP (Boulkenafet, Komulainen, and Hadid 2016; Freitas Pereira et al. 2012), HOG (Tan et al. 2010), and motion patterns (Kollreider, Fronthaler, and Bigun 2009; Bharadwaj et al. 2013), are utilized to highlight distinguishable clues in early attempts. In recent years, deep models have been dominating this community. Diverse network structures, including Convolutional Neural Networks (CNNs) (Yang, Lei, and Li 2014), Recurrent Neural Networks (RNNs) (Xu, Li, and Deng 2015), and Transformer (George and Marcel 2021b) are adopted to learn spoof-relevant spatial and/or temporal features. Additionally, a series of auxiliary tasks (Liu, Jourabloo, and Liu 2018; Atoum et al. 2017; Yu et al. 2020a) are presented for improved performance. Unfortunately, these methods are elaborately designed for a certain type of attack, making them not so competent for more complicated cases, *e.g.*, a wide variety of spoof types co-existing (Liu et al. 2019b). On the other side, generative adversarial deep learning is investigated, and several disentanglement networks are built, which aim to explicitly extract the most fundamental cues in spoofing, *i.e.*, spoof traces, from input faces (Liu, Stehouwer, and Liu 2020; Liu and Liu 2020). They exhibit great potential for coping with both seen and unseen spoof scenarios. Besides, the disentangled spoof traces also reinforce the interpretability of deep FAS models.

Indeed, it is always desirable that sufficient spoofing clues can be captured by RGB cameras. However, due to the intrinsic limitations of the visible spectrum, it is intractable to detect certain types of attacks only by texture analysis, and sophisticated spoofing techniques, *e.g.,* high-quality replayed videos and high-fidelity 3D masks, make this issue even severer. Considering that multi-modal data convey richer information of faces, a number of efforts are made to reinforce FAS by capturing more comprehensive features, where hyper-spectral (Yi et al. 2014), thermal (George et al. 2019), near-infrared (Liu et al. 2021) and depth (Zhang et al. 2020b) images are exploited. With more comprehensive description, these methods prove superior to the ones based on single-modal input (George and Marcel 2021a, 2020;

[*] Corresponding author.

George et al. 2019). This fact suggests a promising alternative to guard FR systems from spoofing attacks in particular for more practical scenarios, that a versatile FAS model can be constructed by leveraging disentanglement learning as well as multi-modal analysis in a unified network. Nevertheless, current multi-modal studies primarily formulate FAS as a classification problem, where spoof-irrelevant cues are inevitably harvested, and this tends to limit the generalization capability.

In this study, we propose a novel approach to FAS, which integrates the advantages of disentanglement networks and multi-modal fusion and collaboratively learns polysemantic spoofing characteristics under a generative framework. First, we choose RGB-D modalities as input, since they provide crucial textures and geometries and the devices are relatively affordable and easy to deploy. Then, based on the adversarial learning mechanism, a two-stream disentanglement model is designed to estimate spoof patterns from the RGB and depth modalities, respectively. In this case, it captures complementary spoofing clues inhering in different attacks. Furthermore, instead of directly concatenating the representations learned in the two modalities, an attention-based fusion module is introduced, which recalibrates heterogeneous representations at multiple stages to assist feature disentanglement in each individual modality. It finally performs cross-modality aggregation for a more powerful spoof trace representation. At the decision phase, the spoof patterns decomposed from both the modalities are fed into a vanilla classification network to predict the attack probability score.

The main contributions include: (i) the first multi-modal disentanglement network for FAS, which effectively improves the accuracy and robustness by learning polysemantic spoof traces; (ii) a feature fusion module, which mines helpful clues across modalities to facilitate the disentanglement; and (iii) the state-of-the-art spoofing detection performance on several publicly available benchmarks under both the seen and unseen protocols.

Related Work

Conventional Face Anti-spoofing

The early face anti-spoofing approaches mostly utilize handcrafted features to extract potential color and texture differences caused by spoofing attacks (Boulkenafet, Komulainen, and Hadid 2016; Tan et al. 2010; Patel, Han, and Jain 2016). Several studies construct dynamic features to emphasize facial actions like blinking (Pan et al. 2011) and micro movements (Kollreider, Fronthaler, and Bigun 2009; Bharadwaj et al. 2013). In the context of deep learning, CNN- and RNN-based methods are introduced (Yang, Lei, and Li 2014; Xu, Li, and Deng 2015) and a series of auxiliary tasks, including recovering depth (Wang et al. 2020b; Atoum et al. 2017), reflection (Yu et al. 2020a; Kim et al. 2019), and rPPG signals (Liu, Jourabloo, and Liu 2018), are considered as constraints to capture fine-grained textural differences. Besides, specific pre-processing and pixel-level supervision are developed to enhance local representations (de Souza, Papa, and Marana 2018; George and Marcel 2019; Sun et al. 2020). However, these solutions heavily deteriorate in practical scenarios where unseen spoofing attacks appear. To solve this problem, a few generalizable methods are designed to mitigate the image distribution shift by performing domain adaptation (Li et al. 2018; Wang et al. 2019a, 2020a; Shao et al. 2019; Jia et al. 2020). Meanwhile, meta-learning based ones are explored (Chen et al. 2021; Qin et al. 2022; Jia, Zhang, and Shan 2022). Despite that gains are continuously reported, the methods above are largely limited by the single RGB modality.

Multi-Model Face Anti-spoofing

(Wang et al. 2018) jointly employs the texture features and the geometric ones extracted from reconstructed depth maps to perform FAS on 3D masks. (Zhang et al. 2019b) proposes a multi-branch network with a squeeze and excitation fusion model to combine RGB, depth, and near-infrared features. (Parkin and Grinchuk 2019) further introduces a multi-level fusion branch for improved results. (Chen et al. 2020a) presents MM-DFN to model multi-modal dynamic features. (Wang et al. 2019b) adopts CBAM (Woo et al. 2018) to build modality-specific liveness features. (Shen, Huang, and Tong 2019) takes patch-level input and designs a modal feature erasing operation to prevent over-fitting. (George and Marcel 2021a) makes use of a cross-modal focal loss to adjust the contribution of each modality while suppressing the impact of uncertain samples. Additionally, input-level and decision-level fusion schemes are considered in (Nikisins, George, and Marcel 2019; Zhang et al. 2019a). However, these multi-modal methods have the difficulty in generalizing to unseen spoofing attacks and along with the conventional ones, their results are hard to interpret.

Generative Face Anti-spoofing

A few disentanglement frameworks are proposed to explicitly decompose spoof patterns from input faces. (Jourabloo, Liu, and Liu 2018) rephrases FAS as a noise modeling problem and establishes an encoder-decoder structure to estimate spoof noise. (Liu, Stehouwer, and Liu 2020; Liu and Liu 2020; Xu et al. 2021) then construct a series of spoof trace elements by adversarial learning. (Feng et al. 2020) applies the classification and metric loss to generate spoof cues in an anomaly detection manner. (Xu et al. 2021) exploits identity information to constrain the estimation of the live face components and generates spoof noise for 2D attacks. Apart from directly generating spoof patterns, several studies use the latent code to disentangle spoof and content features (Zhang et al. 2020a; Wang, Wang, and Lai 2022; Wu et al. 2022). Nevertheless, the previous methods mostly focus on dealing with 2D attacks in the RGB modality. In this study, we leverage the disentanglement mechanism and substantially extend it to the multi-modal space, where we specially investigate the interactions between modalities and present a network to learn polysemantic spoof traces in a collaborative manner. As in (George and Marcel 2021a), RGB-D data are used in this study, but our model can be conveniently adapted to more modalities.

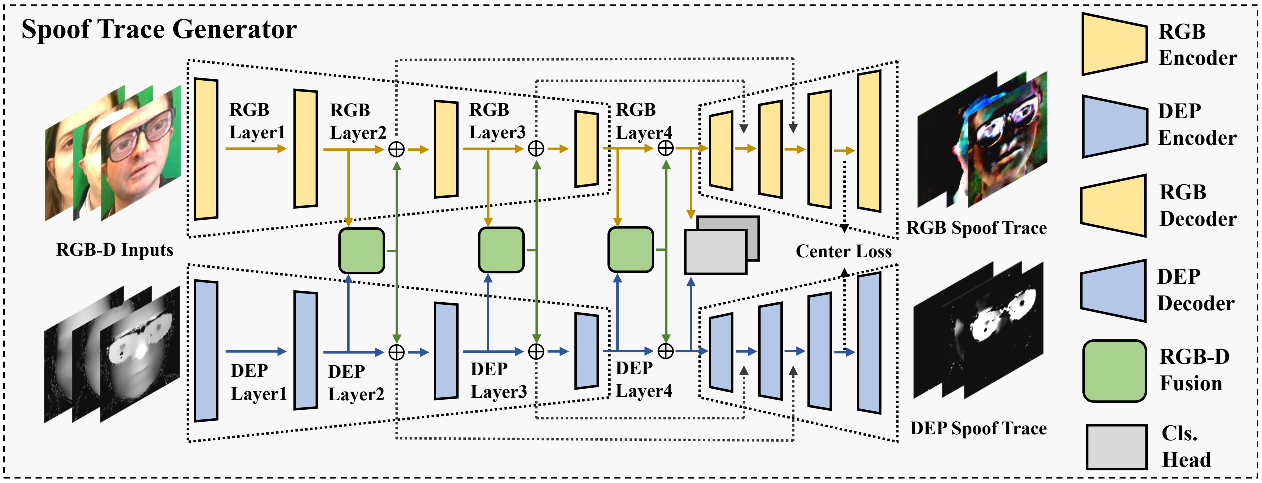

Figure 1: Architecture of the spoof trace generator. It consists of two sub-networks to explicitly disentangle polysemantic spoof traces from the RGB-D inputs. The intermediate features are recalibrated by a cross-modality fusion module at multi-scales.

Method

Our model performs FAS on RGB-D images, which are biased towards textural and geometric spoof clues, respectively. As illustrated in Figure 1, the model is implemented as a generative network, where the two branches explicitly disentangle polysemantic spoof traces from different modalities. To collaboratively learn spoof representations, the intermediate features are recalibrated by a cross-modality fusion module at multi-scales, which not only assists to propagate complementary information between modalities but also helps to preserve modality-specific information. A series of training losses are exploited to guide the training process for improved accuracy and robustness. Both types of the estimated spoof traces are fed into a classification network to make final decision.

Spoof Trace Modeling

Motivated by (Liu, Stehouwer, and Liu 2020; Xu et al. 2021), we regard spoof trace disentangling as a generative problem, where a face can be decomposed into a spoof trace, *i.e.*, the detectable artifact introduced by attacking, and its live counterpart. We uniformly exploit a denoise model (Jourabloo, Liu, and Liu 2018), ensuring that it can be conveniently generalized. For any single modality, an input face image I can be formulated as:

$$I = \hat{I} + T \tag{1}$$

where $\hat{I}$ refers to the live component and T indicates the spoof trace. Theoretically, the decomposed live components should share the same distribution with the real live samples; therefore we extract a spoof trace by reconstructing its original live part via bidirectional adversarial learning. The spoof trace is expected to fully contain spoof-relevant cues, upon which the attack can be detected.

Spoof Trace Generator

Multiple modalities can reinforce the model. The fine-grained color and texture differences caused by spoof instruments are generally conveyed in RGB images, while depth maps mainly contain corresponding geometry variations. To identify the respective strengths of the two modalities and integrate their complementary representations, the spoof trace generator is implemented as a two-stream network to targetedly learn polysemantic spoof traces. By using a cross-modal fusion module, the generator spotlights the intermediate modality-specific features at multiple stages and unifies the features into more comprehensive representations to deliver the final spoof traces.

For either branch, the encoder is constructed by the first three blocks of the ResNet-50 model pre-trained on ImageNet. It gradually encodes the input RGB or depth image into a latent space capturing the spoof-relevant characteristics. The decomposition is achieved by a CNN-based decoder, which consists of multiple stacked residual blocks and deconvolutional layers, yielding the spoof trace conditioned on the input. In the backbone, skip connections are applied for a high-quality transformation. Besides, the cross-modality fusion module is used. To better guide the learning process of the spoof trace generator, an intermediate classification head is also built to encourage the encoder to focus on the spoof-relevant information. It takes the bottleneck feature of the branch as input and outputs a classification score.

Cross-Modality Feature Fusion

Multi-modal data naturally complements each other, but simply assembling heterogeneous information possibly leads to inferior results. Our fusion strategy is inspired by SA-Gate (Chen et al. 2020b), which first conducts Feature Separation (FS) to select the most informative feature maps of each modality via a cross-modality channel attention, and then conducts Feature Aggregation (FA) by adding the cross-modality features with a spatial-wise gate. We take the

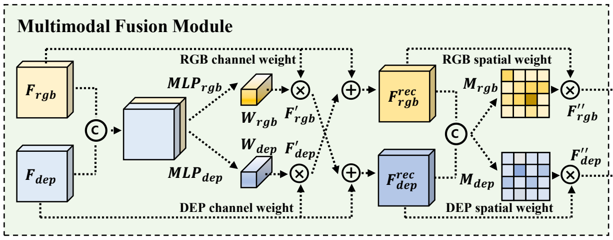

Figure 2: Illustration of the cross-modality fusion module.

advantage of SA-Gate in leveraging both channel-wise and spatial-wise correlations. Unlike the original SA-Gate which produces a single fused representation, we put a higher emphasis on the specificity of each modality, *i.e.*, the RGB and depth information is individually enhanced. In this way, the two kinds of features are then propagated to the next stage of the two-branch generator and their respective spoof traces can be obtained. Figure 2 shows the fusion module.

We take the RGB modality as an example. Let F_{rgb} and F_{dep} denote the learned RGB and depth features. The two types of features are first concatenated, and to filter the most important channels, an MLP is trained to calculate the channel attention between the modalities, followed by a sigmoid function rescaling the weights:

$$W_{rgb} = \sigma(MLP_{rgb}(F_{rgb}||F_{dep})) \tag{2}$$

where $||$ denotes concatenation. A filtered RGB representation, denoted as $F^{'}_{rgb}$, can be obtained via a channel-wise multiplication between the input RGB feature and the cross-modality gate, written as:

$$F^{'}_{rgb} = W_{rgb} \otimes F_{rgb} \tag{3}$$

where $\otimes$ denotes channel-wise multiplication. With the enhanced RGB feature, its depth counterpart can be further recalibrated by borrowing such complementary information from the RGB branch:

$$F^{rec}_{dep} = F^{'}_{rgb} + F_{dep} \tag{4}$$

Similarly, we have $F^{'}_{dep}$ and F^{rec}_{rgb}. The recalibration step is conducted in a symmetric manner so that both modalities can be spotlighted by making them aware of global information. Next, with the refined features, a spatial aggregation step is performed to harvest useful spoof-related context within each modality. In particular, the spatial attention weight is calculated via a 1×1 convolutional layer followed by softmax, which enforces the model to focus on the regions introduced by attacks. Finally, the reinforced feature can be formulated as:

$$\begin{aligned} F^{''}_{rgb} &= F^{rec}_{rgb} \odot M_{rgb}(F^{rec}_{rgb}||F^{rec}_{dep}) \\ F^{''}_{dep} &= F^{rec}_{dep} \odot M_{dep}(F^{rec}_{rgb}||F^{rec}_{dep}) \end{aligned} \tag{5}$$

where M_{rgb} and M_{dep} denote the convolutional operations for the RGB and depth modalities, respectively, and $\odot$ is element-wise multiplication. To fully integrate the two modalities, cross-modality feature fusion is conducted at multiple layers, which enables the complementary information to propagate between modalities throughout the network. With such a network architecture, the RGB and depth features are learned in a mutually reinforced manner, leading to improved disentanglement performance.

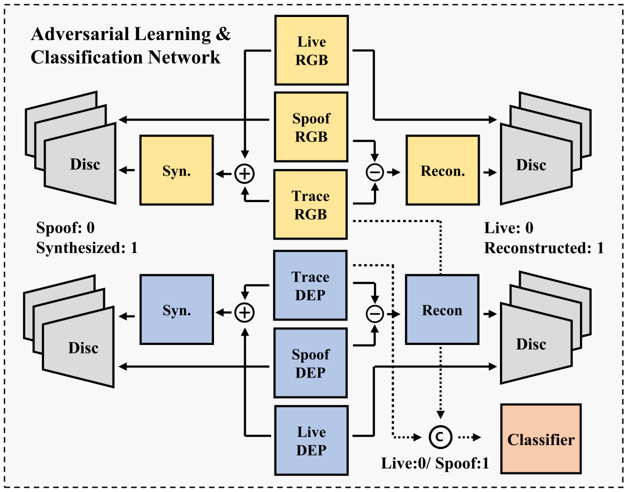

Figure 3: Illustration of the adversarial learning procedure and the final classification network. *Disc* refers to a set of multi-scale discriminators.

Training Process and Loss Functions

Adversarial Loss. Considering that the pixel-level ground-truth labels for spoof patterns are unavailable and the learning process cannot be directly supervised (Jourabloo, Liu, and Liu 2018), we adopt a bidirectional adversarial learning mechanism to achieve spoof trace disentanglement, as illustrated in Figure 3. Specifically, we leverage the following observations: (1) The live components disentangled from the input images and the real face samples are expected to share the same data distribution, and we thus adopt the **general adversarial learning** to encourage the separated live counterparts to be indistinguishable from the real faces. (2) The decomposed spoof patterns can be used to synthesize new spoof images for enhanced disentanglement. For simplicity, we add the obtained spoof pattern to a randomly selected live image to render a synthesized spoof face. In an ideal situation, the synthesized spoof images should share the same distribution with the actual spoof samples. Therefore, the **symmetrical adversarial loss** is calculated.

Motivated by (Liu, Stehouwer, and Liu 2020), multi-scale discriminators are used to estimate the features and the PatchGAN structure (Isola et al. 2017) is adopted, as finer-scale supervision is more suitable for tackling detailed spoof traces. The discriminator is implemented with 7 convolutional layers and 3 down-sampling layers. The adversarial loss for the generator can be formulated as:

$$L_G = \mathbb{E}_{i \in L \cup S}[||D_L(\hat{I}_i)||_2] + \mathbb{E}_{i \in L, j \in S}[||D_S(I_i + T_j)||_2] \tag{6}$$

where D is the multi-resolution discriminator set, and L and S denote the set of live and spoof faces, respectively. T_j is the multi-modal spoof trace disentangled from a spoof sample I_j, and I_i denotes a randomly selected live sample. The discriminators of the two modalities are trained separately. The adversarial loss pushes them to distinguish between the real live samples and the reconstructed ones as well as the actual spoof samples and the synthesized spoof ones, written as:

$$\begin{aligned} L_D =& \mathbb{E}_{i \in L \cup S}[\|D_L(\hat{I}_i) - 1\|_2] + \mathbb{E}_{i \in L}[\|D_L(I_i)\|_2] + \\ & \mathbb{E}_{i \in L, j \in S}[\|D_S(I_i + T_j) - 1\|_2] + \mathbb{E}_{i \in S}[\|D_S(I_i)\|_2] \end{aligned} \tag{7}$$

Identity Consistency Loss. Only using the adversarial loss would cause the decomposed live components to be biased towards the average, and the spoof-irrelevant information, *e.g.*, identity, would be squeezed into the generated spoof trace, especially when the training samples are not so sufficient. To avoid this, an identity consistency loss is utilized to constrain the reconstructed and original input face so that the identity property is preserved. We adopt a lightweight ArcFace model (Deng et al. 2019) based on ResNet-18 as the face recognizer, which measures the identity similarity between the input and reconstructed samples:

$$L_{id} = \mathbb{E}_{i \in L \cup S}[\|f_{Arcface}(I_i) - f_{Arcface}(\hat{I}_i)\|_2] \tag{8}$$

where $f_{Arcface}$ is the pertained face recognition model.

Intensity Loss. Since the real live sample and its live counterpart should be the same, an intensity regularizer is exploited to constrain its spoof trace to be zeros. We also regularize the intensity of the spoof patterns to avoid outliers:

$$L_{intensity} = \mathbb{E}_{i \in L}[\|T_i\|_1] + \lambda_t \mathbb{E}_{i \in S}[\|T_i\|_1] \tag{9}$$

Center Loss. The center loss (Wen et al. 2016) is adopted to optimize the intermediate feature distribution, targeting improved generalization capability to unknown attacks. For each modality, the feature output by the 3rd up-sampling layer of the decoder is used to compute the loss, mainly because this layer is after all the skip-connections thus contains more distinguishable FAS information. The center loss is individually calculated on each modality:

$$L_{center} = \frac{1}{2} \sum_{1}^{2} \|x_i - c_{y_i}\|_2^2 \tag{10}$$

where y_i refers to the ground-truth label which is set to 0 for the live samples and 1 for the spoofs, and c_{y_i} denotes the feature centers of the live and spoof samples, respectively. The calculation of the updated c_{y_i} remains the same as in (Wen et al. 2016).

Classification Loss. The classic cross entropy loss is calculated on the intermediate heads of the two encoders:

$$\begin{aligned} p_i =& \sigma(\psi(E(I_i))) \\ L_e =& \sum_{i=1}^{n} (y_i log(p_i) + (1 - y_i) log(1 - p_i)) \end{aligned} \tag{11}$$

where E is the feature encoder and ψ denotes the classification head of each branch. Besides, the final polysemantic spoof trace is obtained by concatenating the decomposed multi-modal spoof traces in channel, which is fed into a vanilla fully-convolutional network ϕ for classification:

Modality	Method	ACER
RGB-D IR Thermal	MCCNN-BCE	**0.2**
	MCCNN-OCCL-GMM	0.4
	Conv-MLP	0.9
RGB-D	MC-PixBiS	1.8
	MCCNN-OCCL-GMM	3.3
	MC-ResNetDLAS	4.2
	Conv-MLP	6.0
	TTN-T-NHF	0.8
	TTN-S-NHF	0.3
	Ours	**0.27**

Table 1: The evaluation on WMCA, GrandTest protocol.

$$L_{cls} = \mathbb{E}_{i \in L}[\|\phi(T_i)\|_1] + \mathbb{E}_{i \in S}[\|\phi(T_i) - 1\|_1] \tag{12}$$

During training, each mini-batch involves three training steps: generator step, discriminator step, and consistency supervision step. The overall loss of the generator step is:

$$\begin{aligned} L =& \alpha_1 L_G + \alpha_2 L_{intensity} + \alpha_3 L_e + \\ & \alpha_4 L_{id} + \alpha_5 L_{cls} + \alpha_6 L_{center} \end{aligned} \tag{13}$$

At the discriminator step, the model is optimized by $\alpha_7 L_D$, and at the consistency supervision step, the disentangled spoof traces are used to synthesize new spoofs as stated previously. The generator is expected to regenerate these spoof patterns:

$$L = \alpha_8 \mathbb{E}_{i \in L, j \in S}[\|G(I_i + T_j) - T_j\|_2] \tag{14}$$

In order to obtain better generalization ability, the traces of two different spoof samples are also randomly mixed to increase the variety.

Inference. Based on the polysemantic spoof trace, the classification network ϕ predicts the attack probability score.

Experiments

Experimental Settings

Databases and Metrics We conduct extensive experiments on two benchmarks, *i.e.*, WMCA (George et al. 2019) and CASIA-SURF CeFA (Liu et al. 2021). **WMCA** contains seven kinds of 2D and 3D facial presentation attacks, including print, replay, spoof glass, fake head, silicon mask, paper mask, and rigid mask. Multiple modalities are captured in color, depth, infrared, and thermal images. We follow the two major testing protocols on WMCA, *i.e.*, *GrandTest* and *Leave-one-out (LOO)*. **CASIA-SURF CeFA** consists of both 2D and 3D attacks, which additionally considers 3D rigid masks and silicon masks. The data modalities contain RGB, depth, and infrared images. There are four testing protocols: *cross-ethnicity*, *cross-PAI*, *cross-modality* and *cross-ethnicity & PAI*. The 3D attack subset is included only in the evaluation set. In this study, we only focus on the RGB and depth modalities. 5 frames are randomly selected from each video for training and testing. We report APCER (%), BPCER (%), and ACER (%), and the threshold is calculated at BPCER = 1% on the validation set.

Modality	Method	FlexibleMask	Replay	FakeHead	Prints	Glasses	PaperMask	RigidMask	Mean ± Std
RGB	ResNet50	14.5	15.7	38.0	32.7	27.3	20.1	30.2	25.5±9.0
	CDCN	12.1	8.7	42.7	30.1	11.7	11.9	30.4	21.1±13.2
	Auxiliary(Depth)	13.2	12.5	47.3	32.2	23.7	13.9	40.4	26.2±14.1
	TTN-T-NHF	15.1	33.8	1.3	0.4	40.4	3.0	6.0	14.3±16.4
	TTN-S-NHF	**10.7**	21.9	1.3	**0.0**	25.4	**0.0**	2.0	8.8±9.9
RGB-D	MC-PixBiS	49.7	3.7	0.7	0.1	16.0	0.2	3.4	10.5±16.7
	MCCNN-OCCL	22.8	31.4	1.9	30.0	50.0	4.8	18.3	22.7±15.3
	ResNetDLAS	33.3	38.5	49.6	3.8	41.0	47.0	20.6	33.4±14.9
	CMFL	12.4	1.0	2.5	0.7	33.5	1.8	1.7	7.6±11.2
	TTN-T-NHF	26.4	**0.0**	**0.0**	**0.0**	15.9	1.8	8.0	7.4±10.2
	TTN-S-NHF	21.7	1.7	1.7	0.0	21.3	0.7	2.3	7.1±9.9
	Ours	19.0	0.5	2.3	0.7	**10.0**	0.7	**0.6**	**4.8±6.6**

Table 2: The ACER (%) results achieved on WMCA, LOO protocol.

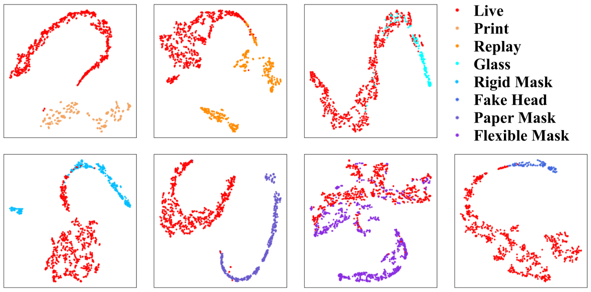

Figure 4: Feature distributions by t-SNE on WMCA, LOO protocol.

Implementation Details Training is launched on a Nvidia Geforce 1080Ti with a batch size of 4 using the Adam optimizer for 40,000 iterations. The initial learning rate is 5e-5. We resize the input face into 256 × 256 pixels. The hyper-parameter λ_t is set to 1e-4, and $\{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7, \alpha_8\}$ are set to $\{0.25, 100, 1, 100, 1, 10, 1, 1\}$.

Quantitative Evaluations

WMCA and CeFA both contain more modalities besides RGB and depth, and some previous studies use all the modalities. To make fair comparison, we reimplement the representative methods by their official codes with default settings, where RGB-D data are adopted for training and testing. For WMCA, the compared methods include MC-PixBiS (Heusch et al. 2020), MCCNN-OCCL-GMM (George and Marcel 2020), MC-ResNetDLAS (Parkin and Grinchuk 2019), CMFL (George and Marcel 2021a), and TTN (Wang et al. 2022b), which signify the-state-of-art. For CeFA, the counterparts include MC-PixBiS, FaceBagNet (Liu et al. 2019a), CDCN (Yu et al. 2020b) and CMFL.

Prot.	Method	ACER	APCER	BPCER
1	MC-PixBiS	15.1±6.9	1.4±0.8	28.7±14.6
	FaceBagNet	17.4±7.9	2.1±1.8	32.7±17.0
	CDCN	6.8±3.5	**0.0±0.0**	13.6±7.0
	MCFL	13.5±9.8	3.7±4.4	23.4±15.6
	Ours	**5.0±1.9**	2.1±3.5	**7.9±5.3**
2	MC-PixBiS	5.9±7.5	10.6±14.9	1.2±0.2
	FaceBagNet	7.9±10.3	15.5±21.0	0.9±0.2
	CDCN	1.2±0.6	**0.0±0.0**	2.5±1.3
	MCFL	2.2±2.6	3.6±5.1	**0.9±0.0**
	Ours	**0.9±0.2**	0.2±0.4	1.5±0.7
4	MC-PixBiS	15.9±1.4	19.0±4.8	12.8±4.6
	FaceBagNet	26.7±9.8	37.9±21.0	15.4±3.2
	CDCN	9.7±6.1	**0.5±1.0**	19.0±12.6
	MCFL	15.2±8.6	11.6±3.4	19.0±21.2
	Ours	**6.2±1.0**	1.3±0.6	**11.1±2.2**

Table 3: The evaluation on CeFA, LOO protocol.

Model	ACER	APCER	BPCER
Baseline	9.0	18.1	0.0
RGB Only	5.7	10.0	1.4
DEP Only	3.8	4.0	3.6
Feature Concat.	1.1	1.5	0.7
Feature Concat. RGB	1.3	2.1	0.6
Feature Concat. DEP	2.5	5.1	**0.0**
SA-Gate	0.52	**0.31**	0.72
Ours	**0.27**	0.39	0.14

Table 4: Ablation study in terms of features and modules on WMCA, GrandTest protocol.

WMCA *GrandTest Protocol:* This evaluation validates the ability to simultaneously detect multiple spoof types. The results are shown in Table 1, where the ones of the compared methods are directly quoted from the original papers. Our approach achieves the best performance compared with the others on RGB-D data, reducing the ACER by 0.03% - 5.73%. The performance is comparable to or even better than that of the methods which exploit more modalities (including MCCNN-BCE (George et al. 2019), MCCNN-OCCL-GMM (George and Marcel 2020), and Conv-MLP (Wang et al. 2022a)), clearly indicating its effectiveness in complex spoofing scenarios.

LOO Protocol: This evaluation aims to test the generalizability in detecting unseen attacks. As Table 2 shows, our model achieves the best overall performance and the individual scores are relatively high for most unseen spoof types. Specifically, promising results are reached on Print, Replay, FakeHead and PaperMask. We also notice that, it does not behave so well for unseen spoofing glasses and sil-

Loss	Mean ± Std (%)
w.o. L_{center}	12.4±13.6
w.o. L_{id}	7.9±9.3
w.o. $L_{intensity}$	6.8±7.3
Ours	**4.8± 6.6**

Table 5: Ablation study in terms of losses on WMCA, LOO protocol.

icon masks (FlexibleMask) although it is still leading by a large margin. The major reason for the failure on glasses is that the training attack signatures cover the entire faces, while their spoof traces only exist in local facial areas with most regions possessing a higher similarity with the real faces. For silicon masks, their texture and geometric properties are inherently quite similar to the real faces, making them more challenging to detect. Overall, our model exhibits a plausible generalizability, which is also evidenced by the t-SNE visualization in Figure 4. As we can see, the real and unseen attack samples reside in obviously distinct regions in the latent space. It confirms that distinguishable spoof traces are disentangled.

CeFA This evaluation is conducted under Protocol 1, 2, and 4 (the cross-modality protocol is not applied). As Table 3 shows, the proposed model achieves competitive spoof detection results for all protocols by only using RGB-D data. In particular, 3D attacks are only included in the test set, which further demonstrates that our method has the favorable robustness and generalizability.

Ablation Study

Ablation is made on WMCA under the GrandTest protocol. Table 4 shows the results, where we compare the following: (1) **Baseline**: no disentanglement is performed. The classification network is trained and tested with the original RGB-D images; (2, 3) **RGB Only** & **DEP Only**: The two branches of the disentanglement network are separately modeled without cross-modality feature fusion. The decision is made upon only one kind of spoof traces; (4, 5, 6) **Feature Concat.** & **Feature Concat. RGB** & **Feature Concat. DEP**: the cross-modality feature fusion is implemented by a vanilla concatenation, and spoof traces of different modalities are used for classification. (7) **SA-Gate**: our feature fusion module is replaced by the original version of SA-Gate.

We can observe: (i) Compared with Baseline, the disentanglement operation significantly improves the performance. Even if only one modality is used, better results are obtained (ACER 5.7% of RGB Only *v.s.* 3.8% of DEP Only *v.s.* 9.0% of Baseline). (2) By comparing the entire model with **RGB Only** and **DEP Only**, the advantage of multi-modal data analysis is validated. A similar conclusion can be drawn from the comparison among (4, 5, 6); (iii) Compared with **Feature Concat.**, the proposed cross-modality fusion module improves ACER by 0.83%, which highlights its contribution in reinforcing the correlations between two modalities. By capturing more complementary information, the disentanglement performance is promoted. (iv) Compared with **SA-gate**, our model improves ACER by 0.25%, further validating that our fusion strategy is more effective.

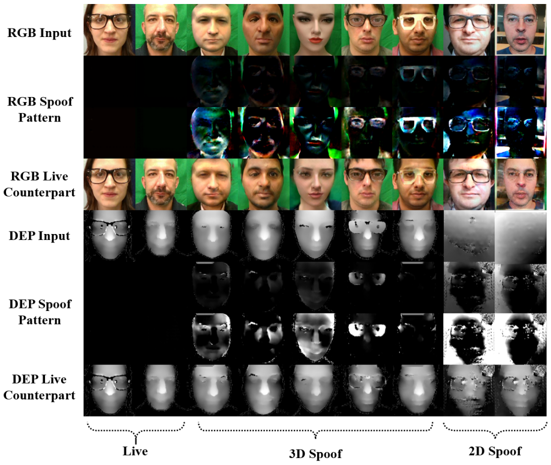

Figure 5: Visualization of the disentangled spoof traces on WMCA.

In Table 5, we perform an ablation under the unseen protocol to verify the necessity of the **training losses**, where the mean ACER over all spoof types is reported. The successively decreasing ACER values demonstrate the effectiveness of the constraints.

Spoof Trace Visualization

The disentangled spoof traces and live counterparts are visualized in Figure 5, with the original inputs shown in the 1st and 5th rows. The 2nd and 6th rows are the extracted spoof traces, which are re-normalized for better view, as displayed in the 3rd and 7th rows. As it can be seen, polysemantic spoof traces are successfully disentangled. Specifically, the RGB modality precisely captures the global color shift and unrealistic local patterns of 3D masks, *e.g.*, shade of eyes, false boundaries, and texture differences (column 3-5). Even without pixel-wise supervision, the local differences around eyes can be detected (column 6-7). The spoof traces of 2D attacks are also well extracted, including the overexposure and color anomalies caused by photo recapturing (column 8-9). Comparatively, the depth modality focuses more on the anomalous geometric changes caused by spoofing, especially for 2D attacks.

Conclusion

This paper presents a novel multi-modal disentanglement network for FAS. By leveraging adversarial disentangled representation learning and cross-modality feature fusion, it generates complementary polysemantic spoof traces from RGB and depth images, respectively, contributing to detecting unseen attacks in more complex scenarios. In particular, it is demonstrated that multi-modal clues are beneficial to facilitating the disentanglement in each single modality. Extensive experiments are conducted on multiple databases and state-of-the-art results are reported.

Acknowledgments

This work is partly supported by the National Natural Science Foundation of China (No. 62176012, 62202031, 62022011), the Beijing Natural Science Foundation (No. 4222049), the Research Program of State Key Laboratory of Software Development Environment (SKLSDE-2021ZX-04), and the Fundamental Research Funds for the Central Universities. We hereby give special thanks to Alibaba Group for their contribution to this paper.

References

Atoum, Y.; Liu, Y.; Jourabloo, A.; and Liu, X. 2017. Face anti-spoofing using patch and depth-based CNNs. In *IJCB*, 319–328.

Bharadwaj, S.; Dhamecha, T. I.; Vatsa, M.; and Singh, R. 2013. Computationally efficient face spoofing detection with motion magnification. In *CVPRW*.

Boulkenafet, Z.; Komulainen, J.; and Hadid, A. 2016. Face spoofing detection using colour texture analysis. *IEEE TIFS*, 11(8): 1818–1830.

Chen, S.; Li, W.; Yang, H.; Huang, D.; and Wang, Y. 2020a. 3d face mask anti-spoofing via deep fusion of dynamic texture and shape clues. In *FG*, 314–321. IEEE.

Chen, X.; Lin, K.-Y.; Wang, J.; Wu, W.; Qian, C.; Li, H.; and Zeng, G. 2020b. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation. In *ECCV*, 561–577.

Chen, Z.; Yao, T.; Sheng, K.; Ding, S.; Tai, Y.; Li, J.; Huang, F.; and Jin, X. 2021. Generalizable representation learning for mixture domain face anti-spoofing. In *AAAI*, volume 35, 1132–1139.

de Souza, G. B.; Papa, J. P.; and Marana, A. N. 2018. On the learning of deep local features for robust face spoofing detection. In *SIBGRAPI*, 258–265.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 4690–4699.

Feng, H.; Hong, Z.; Yue, H.; Chen, Y.; Wang, K.; Han, J.; Liu, J.; and Ding, E. 2020. Learning Generalized Spoof Cues for Face Anti-spoofing. arXiv:2005.03922.

Freitas Pereira, T. d.; Anjos, A.; Martino, J. M. D.; and Marcel, S. 2012. LBP- TOP based countermeasure against face spoofing attacks. In *ACCV*, 121–132.

George, A.; and Marcel, S. 2019. Deep pixel-wise binary supervision for face presentation attack detection. In *ICB*, 1–8.

George, A.; and Marcel, S. 2020. Learning one class representations for face presentation attack detection using multi-channel convolutional neural networks. *IEEE TIFS*, 16: 361–375.

George, A.; and Marcel, S. 2021a. Cross modal focal loss for rgbd face anti-spoofing. In *CVPR*, 7882–7891.

George, A.; and Marcel, S. 2021b. On the effectiveness of vision transformers for zero-shot face anti-spoofing. In *IJCB*, 1–8.

George, A.; Mostaani, Z.; Geissenbuhler, D.; Nikisins, O.; Anjos, A.; and Marcel, S. 2019. Biometric face presentation attack detection with multi-channel convolutional neural network. *IEEE TIFS*, 15: 42–55.

Heusch, G.; George, A.; Geissbühler, D.; Mostaani, Z.; and Marcel, S. 2020. Deep models and shortwave infrared information to detect face presentation attacks. *IEEE TBIOM*, 2(4): 399–409.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *CVPR*, 1125–1134.

Jia, Y.; Zhang, J.; and Shan, S. 2022. Dual-Branch Meta-Learning Network With Distribution Alignment for Face Anti-Spoofing. *IEEE TIFS*, 17: 138–151.

Jia, Y.; Zhang, J.; Shan, S.; and Chen, X. 2020. Single-side domain generalization for face anti-spoofing. In *CVPR*, 8484–8493.

Jourabloo, A.; Liu, Y.; and Liu, X. 2018. Face de-spoofing: Anti-spoofing via noise modeling. In *ECCV*, 290–306.

Kim, T.; Kim, Y.; Kim, I.; and Kim, D. 2019. Basn: Enriching feature representation using bipartite auxiliary supervisions for face anti-spoofing. In *CVPRW*.

Kollreider, K.; Fronthaler, H.; and Bigun, J. 2009. Non-intrusive liveness detection by face images. *IVC*, 27(3): 233–244.

Li, H.; Li, W.; Cao, H.; Wang, S.; Huang, F.; and Kot, A. C. 2018. Unsupervised domain adaptation for face anti-spoofing. *IEEE TIFS*, 13(7): 1794–1809.

Liu, A.; Tan, Z.; Wan, J.; Escalera, S.; Guo, G.; and Li, S. Z. 2021. Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In *WACV*, 1179–1187.

Liu, A.; Wan, J.; Escalera, S.; Jair Escalante, H.; Tan, Z.; Yuan, Q.; Wang, K.; Lin, C.; Guo, G.; Guyon, I.; et al. 2019a. Multi-modal face anti-spoofing attack detection challenge at cvpr2019. In *CVPRW*.

Liu, Y.; Jourabloo, A.; and Liu, X. 2018. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *CVPR*, 389–398.

Liu, Y.; and Liu, X. 2020. Physics-Guided Spoof Trace Disentanglement for Generic Face Anti-Spoofing. arXiv:2012.05185.

Liu, Y.; Stehouwer, J.; Jourabloo, A.; and Liu, X. 2019b. Deep tree learning for zero-shot face anti-spoofing. In *CVPR*, 4680–4689.

Liu, Y.; Stehouwer, J.; and Liu, X. 2020. On disentangling spoof trace for generic face anti-spoofing. In *ECCV*, 406–422.

Nikisins, O.; George, A.; and Marcel, S. 2019. Domain adaptation in multi-channel autoencoder based features for robust face anti-spoofing. In *ICB*, 1–8.

Pan, G.; Sun, L.; Wu, Z.; and Wang, Y. 2011. Monocular camera-based face liveness detection by combining eyeblink and scene context. *Telecommun Syst.*, 47(3): 215–225.

Parkin, A.; and Grinchuk, O. 2019. Recognizing multi-modal face spoofing with face recognition networks. In *CVPRW*.

Patel, K.; Han, H.; and Jain, A. K. 2016. Secure face unlock: Spoof detection on smartphones. *IEEE TIFS*, 11(10): 2268–2283.

Qin, Y.; Yu, Z.; Yan, L.; Wang, Z.; Zhao, C.; and Lei, Z. 2022. Meta-teacher for face anti-spoofing. *IEEE TPAMI*, 44(10): 6311–6326.

Shao, R.; Lan, X.; Li, J.; and Yuen, P. C. 2019. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *CVPR*, 10023–10031.

Shen, T.; Huang, Y.; and Tong, Z. 2019. Facebagnet: Bag-of-local-features model for multi-modal face anti-spoofing. In *CVPRW*.

Sun, W.; Song, Y.; Chen, C.; Huang, J.; and Kot, A. C. 2020. Face spoofing detection based on local ternary label supervision in fully convolutional networks. *IEEE TIFS*, 15: 3181–3196.

Tan, X.; Li, Y.; Liu, J.; and Jiang, L. 2010. Face liveness detection from a single image with sparse low rank bilinear discriminative model. In *ECCV*, 504–517.

Wang, G.; Han, H.; Shan, S.; and Chen, X. 2019a. Improving cross-database face presentation attack detection via adversarial domain adaptation. In *ICB*, 1–8.

Wang, G.; Han, H.; Shan, S.; and Chen, X. 2020a. Unsupervised adversarial domain adaptation for cross-domain face presentation attack detection. *IEEE TIFS*, 16: 56–69.

Wang, G.; Lan, C.; Han, H.; Shan, S.; and Chen, X. 2019b. Multi-modal face presentation attack detection via spatial and channel attentions. In *CVPRW*.

Wang, W.; Wen, F.; Zheng, H.; Ying, R.; and Liu, P. 2022a. Conv-MLP: A Convolution and MLP Mixed Model for Multimodal Face Anti-Spoofing. *IEEE TIFS*, 17: 2284–2297.

Wang, Y.; Chen, S.; Li, W.; Huang, D.; and Wang, Y. 2018. Face anti-spoofing to 3D masks by combining texture and geometry features. In *CCBR*, 399–408.

Wang, Y. C.; Wang, C. Y.; and Lai, S. H. 2022. Disentangled Representation with Dual-stage Feature Learning for Face Anti-spoofing. In *WACV*, 1955–1964.

Wang, Z.; Wang, Q.; Deng, W.; and Guo, G. 2022b. Learning multi-granularity temporal characteristics for face anti-spoofing. *IEEE TIFS*, 17: 1254–1269.

Wang, Z.; Yu, Z.; Zhao, C.; Zhu, X.; Qin, Y.; Zhou, Q.; Zhou, F.; and Lei, Z. 2020b. Deep spatial gradient and temporal depth learning for face anti-spoofing. In *CVPR*, 5042–5051.

Wen, Y.; Zhang, K.; Li, Z.; and Qiao, Y. 2016. A discriminative feature learning approach for deep face recognition. In *ECCV*, 499–515.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *ECCV*, 3–19.

Wu, H.; Zeng, D.; Hu, Y.; Shi, H.; and Mei, T. 2022. Dual Spoof Disentanglement Generation for Face Anti-spoofing with Depth Uncertainty Learning. *IEEE TCSVT*, 32(7): 4626–4638.

Xu, Y.; Wu, L.; Jian, M.; Zheng, W.-S.; Ma, Y.; and Wang, Z. 2021. Identity-constrained noise modeling with metric learning for face anti-spoofing. *Neurocomputing*, 434: 149–164.

Xu, Z.; Li, S.; and Deng, W. 2015. Learning temporal features using LSTM-CNN architecture for face anti-spoofing. In *ACPR*, 141–145.

Yang, J.; Lei, Z.; and Li, S. Z. 2014. Learn Convolutional Neural Network for Face Anti-Spoofing. arXiv:1408.5601.

Yi, D.; Lei, Z.; Zhang, Z.; and Li, S. Z. 2014. Face anti-spoofing: Multi-spectral approach. In *Handbook of Biometric Anti-Spoofing*, 83–102. Springer.

Yu, Z.; Li, X.; Niu, X.; Shi, J.; and Zhao, G. 2020a. Face anti-spoofing with human material perception. In *ECCV*, 557–575.

Yu, Z.; Qin, Y.; Li, X.; Wang, Z.; Zhao, C.; Lei, Z.; and Zhao, G. 2020b. Multi-modal face anti-spoofing based on central difference networks. In *CVPRW*.

Zhang, K.-Y.; Yao, T.; Zhang, J.; Tai, Y.; Ding, S.; Li, J.; Huang, F.; Song, H.; and Ma, L. 2020a. Face anti-spoofing via disentangled representation learning. In *ECCV*, 641–657.

Zhang, P.; Zou, F.; Wu, Z.; Dai, N.; Mark, S.; Fu, M.; Zhao, J.; and Li, K. 2019a. FeatherNets: Convolutional neural networks as light as feather for face anti-spoofing. In *CVPRW*.

Zhang, S.; Liu, A.; Wan, J.; Liang, Y.; Guo, G.; Escalera, S.; Escalante, H. J.; and Li, S. Z. 2020b. Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing. *IEEE TBIOM*, 2(2): 182–193.

Zhang, S.; Wang, X.; Liu, A.; Zhao, C.; Wan, J.; Escalera, S.; Shi, H.; Wang, Z.; and Li, S. Z. 2019b. A dataset and benchmark for large-scale multi-modal face anti-spoofing. In *CVPR*, 919–928.