Towards Generalized Face Anti-Spoofing from a Frequency Shortcut View — Supplementary Material —

Junyi Cao Chao Ma*
MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University

In this supplementary material, Section 1 provides more implementation details of the proposed method. Section 2 introduces the definition of the Intra-class correlation (ICC) mentioned in Section 4.4 of the manuscript. Section 3 displays additional analysis to better demonstrate the effectiveness of our method. Section 4 supplements the comparison results on the OULU-NPU protocol [1]. In Section 5, we display more visual results to facilitate an intuitive understanding of the proposed method. Finally, we analyze the limitations of the proposed method in Section 6.

1. More Implementation Details

In our experiments, we use the MTCNN [19] algorithm for face detection and then adopt a conservative crop that enlarges the facial region by a factor of 1.3 around the center of the tracked face. The cropped facial images are normalized to $[-1, 1]$ and then sent to the framework. Random resized crops and random horizontal flips are used for data augmentation. We implement our method using the PyTorch [11] framework and conduct experiments on a GeForce RTX 2080 Ti GPU. For hyperparameter settings, we fix the patch factor $P = 16$ and the number of chosen masks $K = 4$ mentioned in Section 3.2 and Section 3.3 of the manuscript, respectively.

2. Definition of the Intra-class Corrleation

In this section, we give the definition of the Intra-class correlation (ICC) [7, 9] in detail. Denote $\mathcal{E}$ as a feature extractor and $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \cdots \mathcal{D}_M$ be a dataset with M different categories, where $\mathcal{D}_m = \{(\boldsymbol{X}_i, y_i) | y_i = m\}$ consists of all data in the m-th class. Let $\boldsymbol{f}_i$ be the normalized feature vector of the sample $\boldsymbol{X}_i$ extracted by $\mathcal{E}$, which is defined as:

$$\boldsymbol{f}_i = \frac{\mathcal{E}(\boldsymbol{X}_i)}{\|\mathcal{E}(\boldsymbol{X}_i)\|_2}. \tag{1}$$

Then, the center of the image features from m-th class can be computed by

$$\boldsymbol{\mu}(\mathcal{E}|\mathcal{D}_m) = |\mathcal{D}_m|^{-1} \sum_{\boldsymbol{X}_i \in \mathcal{D}_m} \boldsymbol{f}_i, \tag{2}$$

where $|\mathcal{D}_m|$ denotes the cardinality of $\mathcal{D}_m$. Based on this, the classical intra-class and inter-class variation on the complete dataset $\mathcal{D}$ are defined as

$$V_{\text{intra}}(\mathcal{E}|\mathcal{D}) = \frac{1}{M} \sum_{m=1}^{M} \left(|\mathcal{D}_m|^{-1} \sum_{\boldsymbol{X}_i \in \mathcal{D}_m} \|\boldsymbol{f}_i - \boldsymbol{\mu}(\mathcal{D}_m)\|^2 \right), \tag{3}$$

$$V_{\text{inter}}(\mathcal{E}|\mathcal{D}) = \frac{1}{M(M-1)} \sum_{m=1}^{M} \sum_{j \neq m} \|\boldsymbol{\mu}(\mathcal{D}_j) - \boldsymbol{\mu}(\mathcal{D}_m)\|^2. \tag{4}$$

From the above equations, it is easy to check that V_{intra} measures the feature variation within a specific category, while V_{inter} measures the average pairwise distances of class centers. Following [7], the intra-class correlation (ICC) writes

$$ICC(\mathcal{E}|\mathcal{D}) = \frac{V_{\text{inter}}}{V_{\text{intra}}}. \tag{5}$$

Hence, the ICC value of a feature extractor $\mathcal{E}$ on a dataset $\mathcal{D}$ is larger when the inter-class variation is larger and the intra-class variation is smaller. To this end, the ICC value could measure the discriminability of a feature extractor, given that a good feature embedding has a smaller within-class variation and a larger margin across categories.

3. Additional Analysis

3.1. Ablation study on the proposed constraints

Here, we display additional ablation studies to investigate the effectiveness of the proposed constraints used in our method. We adopt the domain generalization protocol [5, 13, 16] to conduct experiments in this section and show the results in Tab. 1. We first replace the proposed fre-

[*] Corresponding author.

Variants	O&C&I to M		O&M&I to C		O&C&M to I		I&C&M to O		Avg.	
	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC
$\mathcal{L}_{\mathrm{FR}} \rightarrow \mathcal{L}_{\mathrm{IR}}$	6.43	97.99	9.33	96.87	7.85	97.80	**10.80**	95.94	8.60	97.15
w/o $\mathcal{L}_{\mathrm{ST}}$	8.33	97.71	9.22	96.13	9.45	96.61	12.31	94.83	9.83	96.32
Ours	**5.95**	**98.52**	**7.33**	**97.86**	**5.45**	**98.77**	10.88	**96.29**	**7.40**	**97.86**

Table 1. Ablation study on the proposed constraints.

Variants	O&C&I to M		O&M&I to C		O&C&M to I		I&C&M to O		Avg.	
	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC
In. Color & Rec. Color	7.66	97.22	7.74	97.65	10.89	96.69	12.61	94.17	9.73	96.43
In. Color & Rec. Gray (**Ours**)	**5.95**	**98.52**	**7.33**	**97.86**	**5.45**	**98.77**	**10.88**	**96.29**	**7.40**	**97.86**
In. Gray & Rec. Color	11.76	93.61	11.10	94.87	15.50	91.78	13.03	93.11	12.85	93.34
In. Gray & Rec. Gray	11.43	94.18	9.33	94.20	15.45	91.75	12.95	93.05	12.29	93.30

Table 2. Analysis of the frequency-aware autoencoder's input and reconstruction types.

quency reconstruction loss $\mathcal{L}_{\mathrm{FR}}$ with an image reconstruction loss $\mathcal{L}_{\mathrm{IR}}$:

$$\mathcal{L}_{\mathrm{IR}} = \sum_{i}^{P} \sum_{j}^{P} \boldsymbol{W}'_{i,j} \otimes |\mathcal{P}_{i,j}(\hat{\boldsymbol{X}}) - \mathcal{P}_{i,j}(\boldsymbol{X})|, \qquad (6)$$

where $\boldsymbol{W}'_{i,j}$ has a similar definition to $\boldsymbol{W}_{i,j}$ in Eq. (2) of the manuscript but in the spatial domain. As observed, this variant yields a decrease of 0.71% in AUC on average. This suggests that representations achieved by frequency reconstruction are more suitable for face anti-spoofing than those achieved by pixel reconstruction. Then, we ablate the use of the soft-margin triplet loss $\mathcal{L}_{\mathrm{ST}}$ defined in Eq. (3) of the manuscript. This variant yields an average AUC of 96.32%, which is lower than our proposed method, *i.e.*, 97.86%. The quantitative results indicate that the soft-margin triplet loss improves the quality of the latent features for generalizable face anti-spoofing.

3.2. Analysis of the input and reconstruction types

Recall that our frequency-aware autoencoder, trained to restore the frequency features of grayscale images, retains frequency details and gradually filters out stylized features from input color images to obtain intermediate features that are robust to style shifts (refer to Section 3.2 of the manuscript). Here, we experiment with some other combinations of the input and reconstruction types to demonstrate the effectiveness of the proposed design. We present the results in Tab. 2. From the table, it is observed that our choice gets the best result in all settings. As shown in the first row, when the reconstruction target is changed to (frequency representation of) the color image, the variant undergoes obvious performance drops. We suspect that reconstructing RGB frequencies leads to the undesirable preservation of much style information, which is tied to domain properties and negatively influences generalization. Besides, when the inputs to the autoencoder are shifted to grayscale images, the resulting variants fail to leverage incomplete input information for face anti-spoofing, which validates the necessity of using complete representations for the inputs. In summary, these results verify that our design facilitates the retention of style-irrelevant frequency details while avoiding neglecting useful features from complete input information.

Prot.	Methods	APCER(%)	BPCER(%)	ACER(%)
1	FaceDS [6]	1.2	1.7	1.5
	Zhang *et al*. [21]	1.7	0.8	1.3
	Liu *et al*. [10]	0.8	1.3	1.1
	BCN [17]	0.0	1.6	0.8
	CDCN [18]	0.4	1.7	1.0
	DCN [20]	1.3	0.0	0.6
	LMFD-PAD [3]	1.4	1.6	1.5
	PatchNet [15]	0.0	0.0	**0.0**
	Ours	0.4	0.8	0.6
2	FaceDS [6]	3.1	1.9	2.5
	Zhang *et al*. [21]	2.7	2.7	2.4
	Liu *et al*. [10]	2.3	1.6	1.9
	BCN [17]	2.6	0.8	1.7
	CDCN [18]	1.5	1.4	1.5
	DCN [20]	2.2	2.2	2.2
	LMFD-PAD [3]	3.1	0.8	2.0
	PatchNet [15]	1.1	1.2	**1.2**
	Ours	1.7	1.4	1.5
3	FaceDS [6]	2.7±1.3	3.1±1.7	2.9±1.5
	Zhang *et al*. [21]	2.8±2.2	1.7±2.6	2.2±2.2
	Liu *et al*. [10]	1.6±1.6	4.0±5.4	2.8±3.3
	BCN [17]	2.8±2.4	2.3±2.8	2.5±1.1
	CDCN [18]	2.4±1.3	2.2±2.0	2.3±1.4
	DCN [20]	2.3±2.7	1.4±2.6	1.9±1.6
	LMFD-PAD [3]	3.5±3.2	3.3±3.2	3.4±3.1
	PatchNet [15]	1.8±1.5	0.6±1.2	1.2±1.3
	Ours	1.3±0.5	0.0±0.0	**0.6±0.2**
4	FaceDS [6]	5.1±6.3	6.1±5.1	5.6±5.7
	Zhang *et al*. [21]	5.4±2.9	3.3±6.0	4.4±3.0
	Liu *et al*. [10]	2.3±3.6	5.2±5.4	3.8±4.2
	BCN [17]	2.9±4.0	7.5±6.9	5.2±3.7
	CDCN [18]	4.6±4.6	9.2±8.0	6.9±2.9
	DCN [20]	6.7±6.8	0.0±0.0	3.3±3.4
	LMFD-PAD [3]	4.5±5.3	2.5±4.1	3.3±3.1
	PatchNet [15]	2.5±3.8	3.3±3.7	2.9±3.0
	Ours	2.5±2.7	2.5±2.7	**2.5±0.0**

Table 3. Comparison results on the OULU-NPU dataset.

4. Additional Comparison Results

This section includes more comparison results using the OULU-NPU [1] dataset. OULU-NPU consists of four protocols for evaluating the model robustness against unseen environments (*i.e.*, Protocol 1), unseen spoof mediums (*i.e.*, Protocol 2), unseen capture devices (*i.e.*, Protocol 3), and all of the above (*i.e.*, Protocol 4), respectively. The results on these protocols are shown in Tab. 3. It is seen that our approach achieves state-of-the-art results on challenging Protocols 3 and 4 while obtaining comparable performance on Protocols 1 and 2. It is worth noting that previous methods leverage auxiliary supervision like depth maps [6, 18, 21], reflection maps [17], or fine-grained spoof labels [15] to learn discrepancy information, while we only use the binary classification labels. The competitive results yielded by our method demonstrate the effectiveness of the proposed frequency mitigation framework.

5. Additional Visual Results

5.1. Visualization of Classification Decision

In this subsection, we investigate the decision-making mechanism of our proposed method to better understand its effectiveness. Specifically, we display the Grad-CAM [12] visualization of the previous method SSDG [5] and our approach in Figure 1. We can see that SSDG fails to capture the essential discrepancy between live and spoof samples, as it only learns to seek spoofing cues in limited regions. On the contrary, based on the proposed frequency shortcut mitigation, our method produces informative attention maps for separating spoof from live. This discriminative ability mainly results from comprehensive judgment for face anti-spoofing achieved by our framework. For example, regarding the cut paper attacks displayed in the second row of the left side, our method can focus on the paper edge as well as the cut edge near the eyes to detect spoof traces. For the video replay attacks shown in the last row, our method attends to the whole facial region to search for spoof details like Moiré patterns. This visualization, from another perspective, verifies our proposed framework's superiority.

5.2. Visualization of Learning Dynamics

We investigate the learning dynamics of our frequency shortcut mitigation framework to better understand its effectiveness in this subsection. We adopt the I&C&M to O setting from the domain generalization protocol [5, 13, 16] for the following analysis. Regarding image classification tasks, the gradient information $\partial\mathcal{L}/\partial\boldsymbol{X}$ provides us valuable information about the contribution of spatial domain image to classification[1]. Since our framework targets face anti-spoofing through the lens of frequency analysis, we are particularly interested in the gradient information of frequency domain representations. Inspired by previous frequency analysis work [2, 8], we propose to visualize the spectral density of gradients during the training process to study the learning dynamics.

First, we supplement some basic notations used in this subsection. Following the definition in [8], the gradient spectrum of an input sample writes:

$$\boldsymbol{G} = \mathcal{F}(\frac{\partial\mathcal{L}}{\partial\boldsymbol{X}}), \tag{7}$$

where $\mathcal{F}$ indicates the Fast Fourier Transformation (FFT) and $\boldsymbol{G}$ shares the same spatial shape with the input $\boldsymbol{X}$. We denote $\boldsymbol{G}(u, v)$ as a vector located in (u, v) in the gradient spectrum. To investigate which frequency sets attract more attention of the model, we split the gradient spectrum $\boldsymbol{G}$ into several bands $\mathcal{B}_i$ (see Section 3.3 of the manuscript), and then calculate a spectral density scalar S_i for each frequency band. The spectral density measures the azimuthal average of the magnitude of Fourier coefficients over frequencies in a certain band:

$$S_i = |\mathcal{B}_i|^{-1} \sum_{(u,v)\in\mathcal{B}_i} \|\boldsymbol{G}(u, v)\|^2. \tag{8}$$

We use a subset of source domain data to calculate S_i for each frequency band during the first 8,000 training iterations, normalize these values to $[0, 1]$ over all bands at each training iteration, and finally plot the visual results in Figure 2(a)-(b). From the figures, we have the following observations. (1) SSDG [5] roughly focuses on a fixed, limited set of frequency bands during the training process. To detail, the bright regions in Figure 2(a) remain unchanged throughout the training and cover relatively fewer frequency bands compared with our method displayed in Figure 2(b). This indicates that SSDG treats these frequency bands as shortcut representations to simplify the learning process on source domains; however, it may fail on the target domain. (2) On the contrary, our method attends to a broader set of frequency bands as the bright regions shown in Figure 2(b) approximately cover the entire spectrum. Besides, the spectral focus (*i.e.*, the white areas) of our model evolves as the training process continues, promoting the model to attend to under-explored bands for comprehensive judgment.

We further calculate the standard deviation of the normalized spectral density S_i across frequency bands and visualize the trend with respect to training iterations in Figure 2(c). Intuitively, a small standard deviation of S_i across bands means the model allocates relatively even attention to each band, thus alleviating shortcut learning and ensuring a broad focus over the whole frequency spectrum. As shown in this figure, the standard deviation of our model continues to decrease as the training continues, while that of SSDG is

[1] We use $\mathcal{L}$ to denote the cross-entropy loss here for simplicity.

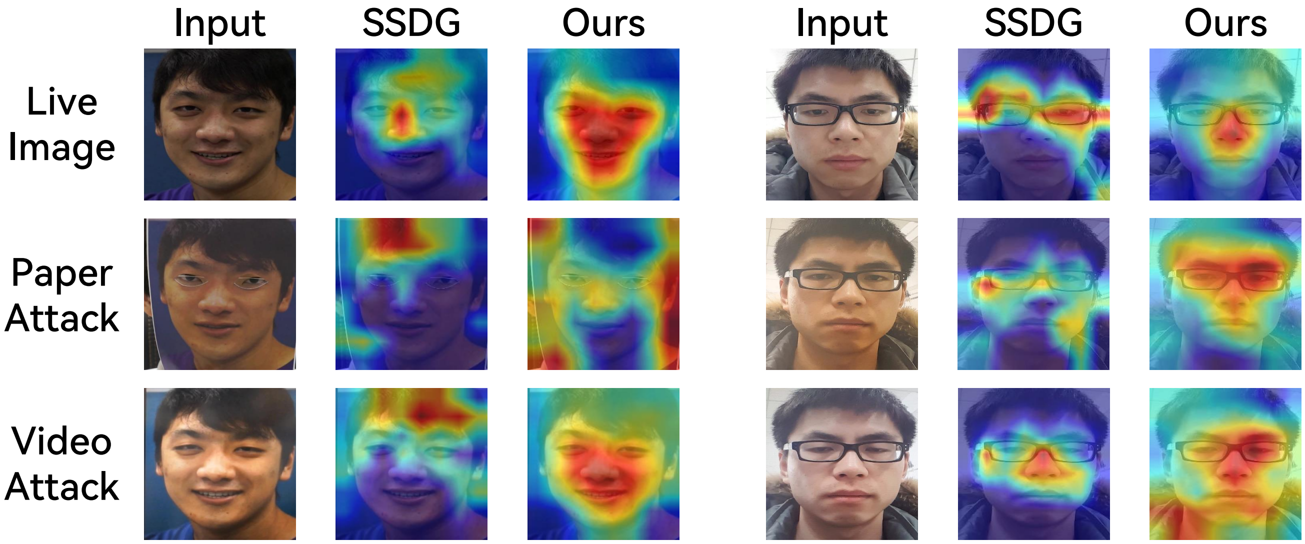
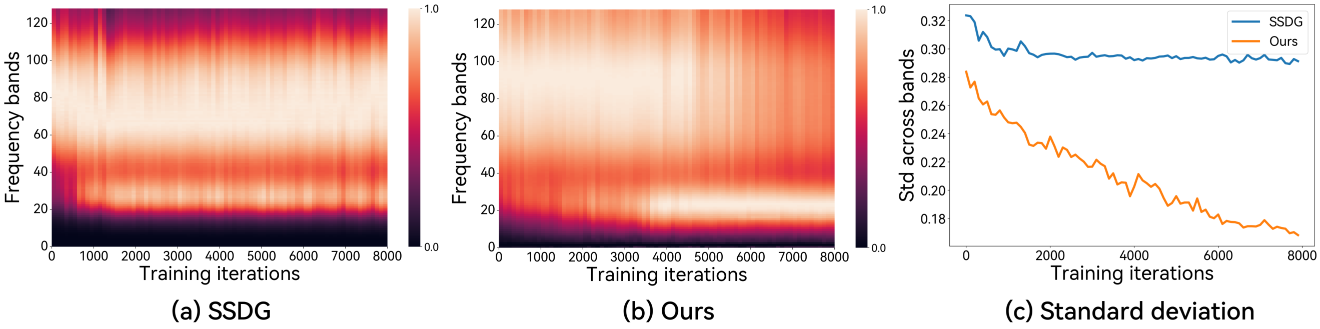


Figure 1. Grad-CAM [12] visualization. Best viewed in color.



(a) SSDG

(b) Ours

(c) Standard deviation

Figure 2. Analysis of the spectral density of gradients during the training process. (a)-(b): Normalized spectral density of gradients at the corresponding frequency bands with respect to the first 8,000 training iterations. (c): Standard deviation of the normalized spectral density across frequency bands measured at specific training iterations.

stuck at a high value. These results effectively verify the superiority of our shortcut mitigation framework, which facilitates the continuous exploration of frequency information for an improved understanding of spoof detection.

6. Limitations

While the proposed method demonstrates its superiority over existing approaches across multiple benchmark datasets, it has limitations. Specifically, our method gears toward shortcut mitigation by broadening the model's attention across a wide frequency range in source domain data. However, it might not adequately generalize to scenarios entirely unfamiliar, where the discrepancies between live faces and spoof images diverge significantly from those encapsulated by the source domain data. Besides, akin to traditional image classifiers, our method remains vulnerable to adversarial attacks [4, 14]. Intricately crafted adversarial inputs could potentially deceive the model, leading to inaccurate predictions.

References

[1] 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. 1, 3

[2] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In *CVPR*, 2020. 3

[3] Meiling Fang, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Learnable multi-level frequency decomposition and hierarchical attention mechanism for generalized face presentation attack detection. In *WACV*, 2022. 2

[4] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversar-

ial examples are not bugs, they are features. In *NeurIPS*, 2019. 4

[5] Yunpei Jia, Jie Zhang, Shiguang Shan, and Xilin Chen. Single-side domain generalization for face anti-spoofing. In *CVPR*, 2020. 1, 3

[6] Amin Jourabloo, Yaojie Liu, and Xiaoming Liu. Face de-spoofing: Anti-spoofing via noise modeling. In *ECCV*, 2018. 2, 3

[7] Hanwen Liang, Qiong Zhang, Peng Dai, and Juwei Lu. Boosting the generalization capability in cross-domain few-shot learning via noise-enhanced supervised autoencoder. In *ICCV*, 2021. 1

[8] Zhiyu Lin, Yifei Gao, and Jitao Sang. Investigating and explaining the frequency bias in image classification. In *IJCAI*, 2022. 3

[9] Bin Liu, Yue Cao, Yutong Lin, Qi Li, Zheng Zhang, Mingsheng Long, and Han Hu. Negative margin matters: Understanding margin in few-shot classification. In *ECCV*, 2020. 1

[10] Yaojie Liu, Joel Stehouwer, and Xiaoming Liu. On disentangling spoof trace for generic face anti-spoofing. In *ECCV*, 2020. 2

[11] Adam Paszke, S. Gross, Soumith Chintala, G. Chanan, E. Yang, Zachary Devito, Zeming Lin, Alban Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017. 1

[12] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *CVPR*, 2017. 3, 4

[13] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *CVPR*, 2019. 1, 3

[14] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014. 4

[15] Chien-Yi Wang, Yu-Ding Lu, Shang-Ta Yang, and Shang-Hong Lai. PatchNet: A simple face anti-spoofing framework via fine-grained patch recognition. In *CVPR*, 2022. 2, 3

[16] Zhuo Wang, Zezheng Wang, Zitong Yu, Weihong Deng, Jiahong Li, Tingting Gao, and Zhongyuan Wang. Domain generalization via shuffled style assembly for face anti-spoofing. In *CVPR*, 2022. 1, 3

[17] Zitong Yu, Xiaobai Li, Xuesong Niu, Jingang Shi, and Guoying Zhao. Face anti-spoofing with human material perception. In *ECCV*, 2020. 2, 3

[18] 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. In *CVPR*, 2020. 2, 3

[19] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 1

[20] Ke-Yue Zhang, Taiping Yao, Jian Zhang, Shice Liu, Bangjie Yin, Shouhong Ding, and Jilin Li. Structure destruction and content combination for face anti-spoofing. In *IJCB*, 2021. 2

[21] Ke-Yue Zhang, Taiping Yao, Jian Zhang, Ying Tai, Shouhong Ding, Jilin Li, Feiyue Huang, Haichuan Song, and Lizhuang Ma. Face anti-spoofing via disentangled representation learning. In *ECCV*, 2020. 2, 3