Multi-modal Face Presentation Attack Detection via Spatial and Channel Attentions

Guoqing Wang[1,3], Chuanxin Lan[1], Hu Han[*,1,2], Shiguang Shan[1,2,3,4], and Xilin Chen[1,3]

[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]University of Chinese Academy of Sciences, Beijing 100049, China
[4]CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai, China
{guoqing.wang, chuanxin.lan}@vipl.ict.ac.cn, {hanhu, sgshan, xlchen}@ict.ac.cn

Abstract

Face presentation attack detection (PAD) has drawn increasing attentions to secure face recognition (FR) systems which are being widely used in many applications from access control to smartphone unlock. Traditional approaches for PAD may lack good generalization capability into new application scenarios due to the limited number of subjects and data modality. In this work, we propose an end-to-end multi-modal fusion approach via spatial and channel attention to improve PAD performance on CASIA-SURF. Specifically, we first build four branches integrated with spatial and channel attention module to obtain the uniform features of different modalities, i.e., RGB, Depth, IR and the fused modality with 9 channels which concatenating three modalities. Subsequently, the features extracted from the four branches are concatenated and fed into the shared layers to learn more discriminative features from the fusion perspective. Finally, we get the classification confidence scores w.r.t. PAD or not. The entire network is optimized with the joint of the center loss and softmax loss and SGRD solver to update the parameters. The proposed approach shows promising results on the CASIA-SURF dataset.

1. Introduction

Face presentation attack detection (PAD) is an important problem in computer vision, which aims to determine whether the captured face is a live or spoof face in the face recognition (FR) systems [29]. It is well known that most of the FR systems are vulnerable to face presentation attack (PA), e.g., print attack, video replay attack, and 2D/3D mask attack, etc. Therefore, face PAD is a very important step of the FR systems and an urgent problem to be solved.

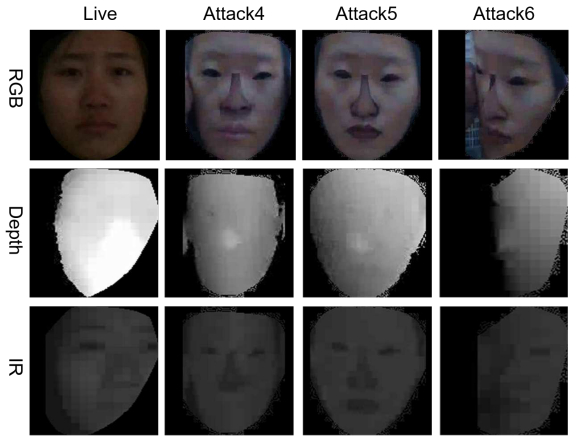

Figure 1. Some examples of the live and spoof faces from the CASIA-SURF.

Some previous face PAD approaches have achieved great performance on 2D presentation attacks, such as print attack and replay attack. These methods assume that there are inherent disparities between live and spoof faces, e.g., skin detail loss, color distortion, moiré pattern, shape deformation, and spoof artifacts, etc. These factors are then utilized to design hand-crafted features for binary classification with a SVM model [3, 7, 13, 22, 31, 28, 27].

Recently, Convolutional Neural Networks (CNNs) have demonstrated its success in many computer vision tasks and a lot of current PAD approaches utilized CNNs for end-to-end face PAD or representation learning followed by bi-

*Corresponding author.

nary classification using SVM [26, 35]. Furthermore, some PAD approaches considered that it is not reasonable to regard the face PAD as merely a binary classification problem and utilized some auxiliary-driven cues such as rPPG signal and depth information to supervise the CNN learning [1, 19, 24, 23]. However, the PAD generalization performance drops significantly under new application scenarios due to the limited number of subjects and data modality. Zhang et al. [36] introduced a large-scale multi-modal face anti-spoofing dataset, namely CASIA-SURF, and make it possible to solve the challenge with a multi-modality perspective. Fig. 1 shows three modal frames of live and 3 different attack ways in training sets.

In this paper, we propose a multi-modal approach to effectively leverage the information in RGB, Depth and IR modalities, which utilizes attention mechanism along channel and spatial dimensions to learn which information is more information and generative for the PAD task. In particular, RGB, Depth, IR and three modalities combined into 9 channels input for ResNet-18 for feature learning with attention mechanism. In order to enhance the discriminative power of the deeply learned features, the network is using SGRD strategy to update the parameters and optimized with the joint supervision of softmax loss and center loss [32], aiming to minimize the intra-class variations while keep the features of different classed separable. Our approach is end-to-end trainable, and achieves promising results in CASIA-SURF dataset.

The main contributions of this work are three-fold: (i) a novel fusion network architecture for multi-modal face PAD with spatial and channel attentions; (ii) SGRD solver to update network parameters and joint supervision of softmax loss and center loss to obtain more discriminative feature representation for live and spoof faces; and (iii) good performance on the CASIA-SURF multi-modal face anti-spoofing dataset.

2. Related Work

2.1. Methods

In the past few years, a number of PAD methods have been proposed, which can be generally divided into hand-crafted feature based methods and deep learning based methods.

1) Hand-crafted feature based methods: Early PAD works utilized hand-crafted features to distinguish between live and spoof faces, such as LBP [7, 22], HoG [13], and SIFT [27]. Some works adopted contextual information [13] to design features. And some other works adopt face motion analysis such as eyes, mouth [25, 12] and 3D geometry analysis [17]. In order to improve the robustness to new scenario, HSV, YCbCr color space [2, 3] and Fourier spectrum space [17] are utilized to get the hand-crafted features.

These hand-crafted features based methods can work well under intra-database testing scenario with low computational complexity. However, the hand-crafted features are intuitively designed based on limited scenario, which have poor generalization ability in cross-database PAD detection scenario.

2) Deep learning based methods: In recent years, a lot of methods [10, 26, 35] based on CNN have emerged, which achieve great success. These methods use CNN-based feature representations or the end-to-end CNN network for binary classification. Yang et al. [35] implemented a canonical CNN structure for learning PAD features. Xu et al. [34] adopted temporal features by combining LSTM and CNN. Liu et al. [19] designed a novel framework to leverage the auxiliary information of depth and rPPG signals in order to learn discriminative and generalizable cues from a face video. Jourabloo et al. [11] inversely decomposed a spoof face into a spoof noise and a live face and then utilized the spoof noise for classification. Wang et al. [30] utilized facial depth for PAD, which is recovered from temporal information. Liu et al. [20] extracted the normal cues via light reflection analysis and then used them to recover subjects' depth maps and also provide the light CAPTCHA checking mechanism to assist liveness classification. In order to improve PAD generalization capability, Li et al. [15] utilized an unsupervised domain adaptation to learn a more generalized classifier.

The deep learning based methods show better performance than the traditional hand-crafted feature based methods under limited scenarios. However, these methods have unsatisfied generalization ability due to the limited number of subjects and data modalities.

2.2. Datasets

Datasets are very important for PAD methods, which directly affect the performance and generalization ability of the model. Most of existing PAD datasets only have RGB modal, such as Replay-Attack [5], CASIA-FASD [37], MSU-MFSD [31], OULU-NPU [4] and SiW [19]. These datasets are captured using several acquisition devices with different resolutions and include multiple attack types, e.g., photo warping attack, cutting attack and replay attack.

With the development of attack technologies, some new types of PA have emerged, such as 3D and silicone masks, which are extremely similar to genuine faces. One way to make the system robust to these attacks is to collect new high-quality databases. Therefore, some datasets include other modality information with the development of sensors. Kose et al. [14] propose a 2D+3D face mask attacks dataset, which is not public. Erdogmus et al. [9] proposed the first publicly available 3D spoofing database (3DMAD),

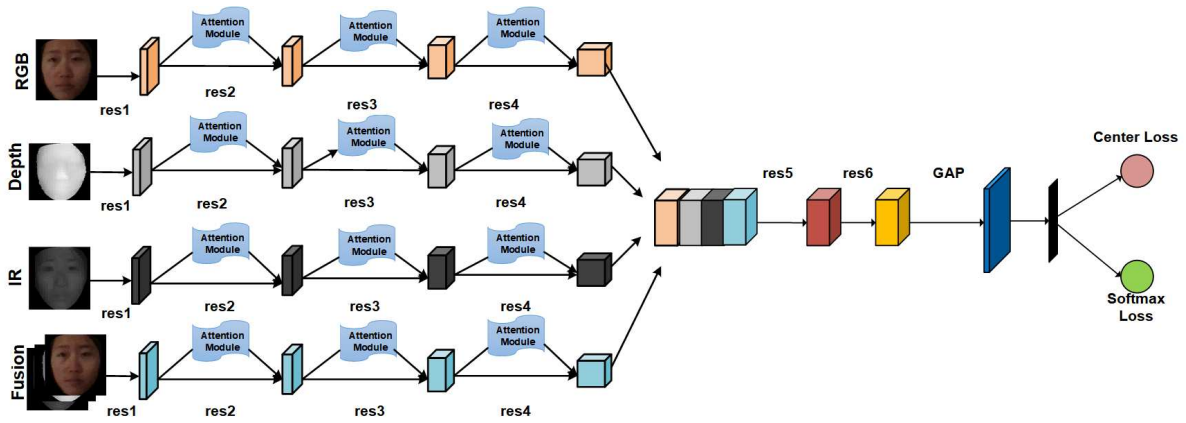

Figure 2. The overall diagram of the proposed approach for multi-modal face presentation attack detection via spatial and channel attention. We first build four branches which both have four residual blocks integrated with the channel and spatial attention module (i.e., res1, res2, res3, res4) to receive RGB, Depth, IR and fusion modal inputs. Subsequently, we concatenate the features extracted from the four branches and fed into the res5 and res6 block which are shared to learn more discriminative features. The entire fusion network is optimized with center loss and softmax loss. GAP means the global average pooling.

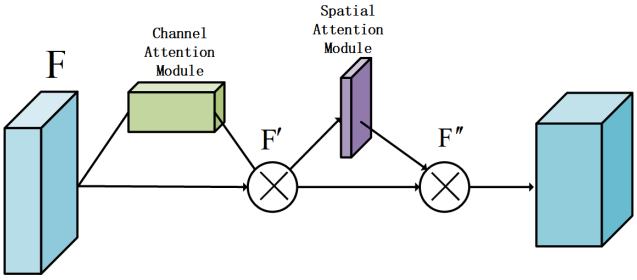

Figure 3. The architecture of the channel and spatial attention module in the proposed PAD method.

recorded with a low-cost depth sensor and show that using color and depth modalities can perform better than only one modality. Another dataset is Multispectral-Spoof [6], which contains VIS and NIR multispectral to reduce the spoofing attacks security risk. The datasets mentioned above have the limited number of subjects and samples, which limit the further research. To solve this problem, Zhang et al. [36, 16] proposed CASIA-SURF, which is a large-scale multi-modal dataset and contains 1,000 subjects with 21,000 videos and each sample has three modalities (i.e., RGB, Depth and IR).

3. Proposed Method

As shown in Fig. 2, our approach consists of four branches, i.e., RGB modality branch, Depth modality branch, IR modality branch and the branch which fuses the above three modalities. The extracted features from these four branches are then concatenated and fed into the shared layers to get the final classification results.

3.1. Attention Fusion

Since the CASIA-SURF dataset is characterized by multi-modal (i.e, RGB, Depth, and IR), the key point is to find a straightforward architecture which can make full use of the complementary information between the three modalities. We build a multi-stream architecture and use the feature-level fusion category which is to fuse the features extracted from RGB, Depth, IR and fused modality subnetworks and then fed into the shared layers to learn joint representations. Each subnetwork has four residual blocks and the channel and spatial attention modules are embedded between two residual blocks inspired by [33]. $F \in \mathbb{R}^{C\times H\times W}$ denotes the input feature map. Subsequently, a 1D channel attention map $M_c \in \mathbb{R}^{C\times 1\times 1}$ and a 2D spatial attention map $M_s \in \mathbb{R}^{1\times H\times W}$ as illustrated in Fig. 3 are inferred through the network training process. The overall attention process can be summarized as:

$$F^{'} = M_c(F) \otimes F, \tag{1}$$

$$F^{''} = M_s(F^{'}) \otimes F^{'}, \tag{2}$$

where $\otimes$ denotes element-wise multiplication. Through channel attention and spatial attention operation, we will get the final refined output $F^{''}$.

3.2. Joint Loss

The choice of loss function directly affects the discriminative power of the deeply learned features. Intuitively, minimizing the intra-class variations while keeping the features of different classes separable is the key goal. So we choose to use center loss and softmax loss to jointly supervise the network training. The joint loss functions are as

Method	TPR (%)			APCER (%)	NPCER (%)	ACER (%)
	@FPR=10^{-2}	@FPR=10^{-3}	@FPR=10^{-4}			
Halfway fusion in [36]	89.1	33.6	17.8	5.6	3.8	4.7
SE fusion in [36]	96.7	81.8	56.8	3.8	1.0	2.4
Three branch fusion	**99.9**	98.7	95.3	0.5	**0.1**	0.3
Four branch fusion	**99.9**	**99.1**	**97.6**	**0.2**	0.3	**0.2**

Table 1. Effectiveness of the proposed fusion method. All models are trained in the CASIA-SURF training set and tested in the testing set.

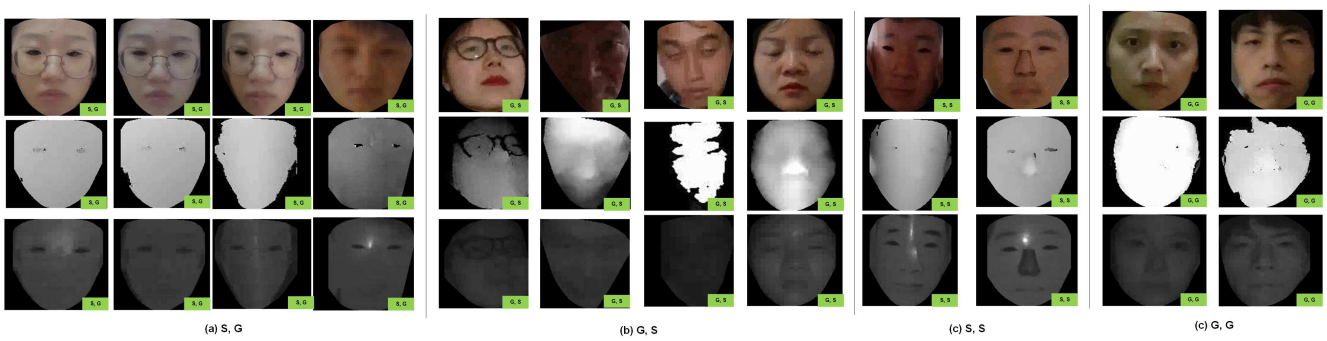

Figure 4. Examples of correct and incorrect PAD results by the proposed approach in CASIA-SURF database tests. The label 'S, G' (or 'G, S') denotes a spoof (genuine) face image is incorrectly classified as genuine (spoof) face image; 'G, G' (or 'S, S') denotes a genuine (spoof) face image is correctly classified as genuine (spoof).

follows:

$$\mathcal{L}_c = \frac{1}{2}\sum_{i=1}^{m} \|x_i - c_{y_i}\|_2^2, \tag{3}$$

$$\mathcal{L}_s = -\sum_{i=1}^{m} \log \frac{e^{\omega_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{\omega_{y_j}^T x_j + b_{y_j}}}, \tag{4}$$

$$\mathcal{L} = \mathcal{L}_c + \lambda \mathcal{L}_s \tag{5}$$

where $c_{y_i} \in \mathbb{R}^d$ denotes the y_ith class center of deep features. The hyperparameter λ is used for balancing the two loss functions, and we impractically set λ = 1 in our experiments below.

3.3. Network Training

Restart techniques can be used when training deep neural networks to obtain averaged gradients, because the gradients can vary significantly from one batch of the data to another. Loshchilov et al. [21] proposed a simple warm restart mechanism, namely stochastic gradient descent with warm restarts (SGDR) to improve the conventional SGD's performance. Specifically, the restarts are not performed for initialling solution but emulated by initialling the learning rate to certain values.

Within the i-th run, the learning rate of SGDR is decayed with a cosine annealing for each batch as follows:

$$\eta_t = \eta_{min}^t + \frac{1}{2}(\eta_{max}^i - \eta_{min}^i)(1 + \cos(\frac{T_{cur}}{T_i}\pi)), \tag{6}$$

where i is the index of the run, T_i denotes how many epochs should be performed within the i-th run. η_{min}^i and η_{max}^i are the ranges for the learning rate, and T_{cur} accounts for how many epochs have been performed since the last restart.

Inspired by [18, 8], we use a general-to-specific transfer learning scheme which first pre-train on the other large-scale PAD datasets, such as, CASIA-FASD [37], MSU-MFSD [31] and Replay-Attack [5] and then we train the whole network jointly using the CASIA-SURF [36].

We resize the cropped face region to the size 56 × 56, and use an open source imgaug [1] library to do data augmentation, i.e., random flipping, rotation, resizing, cropping and color distortion. For the CASIA-SURF dataset [36], our model is trained with end-to-end style for 200 epochs. The model is optimized by the SGDR solver on 4 TITAN XP GPU with a mini-batch 1024. Weight decay and momentum are set to 0.0005 and 0.9, respectively.

4. Experimental Results

4.1. Database and Settings

We provide evaluations on the CASIA-SURF face anti-spoofing dataset [36], which contains multi-modal (RGB, Depth and IR) face images. Specifically, this dataset contains 1,000 Chinese people and each person has 1 live video clip and 6 fake video clips (6 different attack manners) for each modality. These RGB, Depth and Infrared (IR) videos

[1] https://github.com/aleju/imgaug

Method	TPR (%)			APCER (%)	NPCER (%)	ACER (%)
	@FPR=10^{-2}	@FPR=10^{-3}	@FPR=10^{-4}			
RGB modal	89.5	69.5	39.8	5.2	2.6	3.9
Depth modal	99.5	81.5	55.8	0.7	0.8	0.7
IR modal	96.5	64.2	44.2	3.2	**0.2**	1.7
Proposed method	**99.9**	**99.1**	**97.6**	**0.2**	0.3	**0.2**

Table 2. Performance of intra-database testing for individual modalities. All models are trained in the CASIA-SURF training set and tested on the testing set.

Method	TPR (%)			APCER (%)	NPCER (%)	ACER (%)
	@FPR=10^{-2}	@FPR=10^{-3}	@FPR=10^{-4}			
w/o attention	99.1	90.6	83.4	0.6	1.2	0.9
w/o joint loss	99.9	98.7	95.3	0.5	**0.1**	0.3
w/o attention&joint loss	96.7	81.8	56.8	3.8	1.0	2.4
Proposed method	**99.9**	**99.1**	**97.6**	**0.2**	0.3	**0.2**

Table 3. Performance of proposed method under ablation study in terms of attention mechanism and joint loss. All models are trained in the CASIA-SURF training set and tested on the validating set.

simultaneously captured using the Intel RealSense SR300 camera. The background image area of the face was removed from original videos to make the face PAD task more challenging. We split the database into training, validation and testing sets, which contain 300, 100, and 600 subjects and 148K, 48K, and 295K frames, respectively, after selecting one frame out of every 10 frames and removing non-detected face poses with extreme lighting conditions.

We evaluate our method following the intra-database protocol [36], which may be different from the other participants' protocols in the CVPR ChaLearn competition [2], which uses the live faces and attacks no. 4, 5, 6 as the final training and validation sets and uses the live faces and attacks no. 1, 2, 3 as the testing set for the final evaluation. Different attack types are included in the training and testing sets to increase the difficulty of face anti-spoofing detection task. We show some examples of live and spoof face images of three modalities in Fig. 1.

4.2. Comparison with Baselines

We use the methods of [36] as the baseline, which also uses a fusion based approach for PAD, namely halfway fusion and SE fusion, respectively. In addition, we also use design two fusion based approaches as the baselines, namely three branch fusion and four branch fusion. The three branches fusion method use three branch integrated with spatial and channel attention module to extract the features and then concatenate the extracted features and fed into the shared layers optimized with joint loss to improve the intra-database testing performance and report promising results. The four branch fusion method has one more branch to extract the features in fusion modality.

The results are shown in Table 1, from which we can observe that the SE Fusion method in [36] achieves better performance than the native halfway fusion (Half Fusion) in [36], especially at a very low FPR=10^{-3} and 10^{-4}. This suggests that the effectiveness of the squeeze and excitation fusion used in [36]. We also notice that our baseline fusion performs better than [36]. This suggests that the proposed attention fusion method supervised with joint loss is more effective than the proposed squeeze and excitation fusion method in [36]. The proposed method achieves much better PAD performance than the baselines, which shows that the four-branch fusion could obtain more discriminative cues than three branch in the network learning process.

Fig. 4 shows some examples of correct and incorrect PAD results by the proposed approach under intra-database testing. We notice that most errors are caused when the testing face images have appearance variances such as reflective of wearing glasses, dim illumination, similar color distortions in both live and spoof face images, etc.

4.3. Comparison of Multi-modal

Since many previous methods on face PAD reported their performance on only RGB modality due to the limited datasets, we also perform intra-database testing on all modalities to demonstrate they can complement each other.

The results are shown in Table 2, which shows that the depth modal obtains better performance than RGB and IR modalities, especially at FPRs=10^{-3} and 10^{-4}. The IR modality also leads to higher results than the RGB modality. So the introduction of multi-modal PAD in [36] is necessary and is expected to the research of PAD. The results show that our method can better leverage the complemen-

[2] https://competitions.codalab.org/competitions/20853

tary information from different modalities, and achieve better performance.

4.4. Abalation Study

We provide ablation study to validate the two key components in the proposed method: (i) attention mechanism and (ii) joint loss. We study their influences by removing one component each time, and denote the corresponding model as 'w/o attention' and 'w/o joint loss', respectively. The results under intra-database testing are given in Table 3. We can see removing either component can lead to performance drop. This suggests that both components are useful in the proposed face PAD fusion approach.

5. Conslusion

We propose an end-to-end approach for face presentation attack detection (PAD) by mining the complementary information contained in RGB, Depth, and IR using spatial and channel attentions. We first build four branches integrated with spatial and channel attention module to obtain the unique features of different modalities, i.e., RGB, Depth, IR and the fusion modality which concatenates three modalities. Then the extracted features from four branches are concatenated and fed into the shared layers to classify supervised with the joint of the center loss and softmax loss. The proposed approach obtains the promising results on the CASIA-SURF dataset. Our feature work includes utilizing the 3D face prior knowledge and physiological cues to improve the robustness of PAD. In addition, we will also study how to learn better representations that can minimize the influences by subjects identity, race, etc.

6. Acknowledgement

This research was supported in part by the Natural Science Foundation of China (grants 61732004, 61390511, and 61672496), External Cooperation Program of Chinese Academy of Sciences (CAS) (grant GJHZ1843), and Youth Innovation Promotion Association CAS (2018135).

References

[1] Yousef Atoum, Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Face anti-spoofing using patch and depth-based CNNs. In *Proc. IJCB*, pages 319–328, 2017. 2

[2] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face anti-spoofing based on color texture analysis. In *Proc. ICIP*, pages 2636–2640, 2015. 2

[3] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face antispoofing using speeded-up robust features and fisher vector encoding. *IEEE Signal Proc. Let.*, 24(2):141–145, 2017. 1, 2

[4] Zinelabdine Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdenour Hadid. OULU-NPU: A mobile face presentation attack database with real-world variations. In *Proc. FG*, pages 612–618, 2017. 2

[5] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *Proc. BIOSIG*, 2012. 2, 4

[6] Ivana Chingovska, Nesli Erdogmus, André Anjos, and Sébastien Marcel. Face recognition systems under spoofing attacks. In *Face Recognition Across the Imaging Spectrum*, pages 165–194. 2016. 3

[7] Tiago de Freitas Pereira, André Anjos, José Mario De Martino, and Sébastien Marcel. Lbp-top based countermeasure against face spoofing attacks. In *Proc. ACCV*, pages 121–132, 2012. 1, 2

[8] Hui Ding, Shaohua Kevin Zhou, and Rama Chellappa. Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In *Proc. FG*, pages 118–126, 2017. 4

[9] Nesli Erdogmus and Sébastien Marcel. Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect. In *Proc. BTAS*, pages 1–6, 2013. 2

[10] 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. *J. Vis. Commun. Image Represent.*, 38:451–460, 2016. 2

[11] Amin Jourabloo, Yaojie Liu, and Xiaoming Liu. Face de-spoofing: Anti-spoofing via noise modeling. *arXiv preprint, arXiv:1807.09968*, page 3, 2018. 2

[12] Klaus Kollreider, Hartwig Fronthaler, Maycel Isaac Faraj, and Josef Bigun. Real-time face detection and motion analysis with application in liveness assessment. *IEEE Trans. Inf. Forensics Security*, 2(3):548–558, 2007. 2

[13] Jukka Komulainen, Abdenour Hadid, and Matti Pietikainen. Context based face anti-spoofing. In *Proc. BTAS*, pages 1–8, 2013. 1, 2

[14] Neslihan Kose and Jean-Luc Dugelay. Countermeasure for the protection of face recognition systems against mask attacks. In *Proc. FG*, pages 1–6, 2013. 2

[15] Haoliang Li, Wen Li, Hong Cao, Shiqi Wang, Feiyue Huang, and Alex C Kot. Unsupervised domain adaptation for face anti-spoofing. *IEEE Trans. Inf. Forensics Security*, 13(7):1794–1809, 2018. 2

[16] Ajian Liu, Jun Wan, Sergio Escalera, Hugo Jair Escalante, Zichang Tan, Qi Yuan, Kai Wang, Chi Lin, Guodong Guo, Isabelle Guyon, and Stan Z. Li. Multi-modal face anti-spoofing attack detection challenge at CVPR2019. In *Proc. CVPRW*, 2019. 3

[17] Siqi Liu, Pong C Yuen, Shengping Zhang, and Guoying Zhao. 3D mask face anti-spoofing with remote photoplethysmography. In *Proc. ECCV*, pages 85–100, 2016. 2

[18] Xin Liu, Shaoxin Li, Meina Kan, Jie Zhang, Shuzhe Wu, Wenxian Liu, Hu Han, Shiguang Shan, and Xilin Chen. Agenet: Deeply learned regressor and classifier for robust apparent age estimation. In *Proc. ICCVW*, pages 258–266, 2015. 4

[19] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *Proc. CVPR*, pages 389–398, 2018. 2

[20] Yao Liu, Ying Tai, Ji-Lin Li, Shouhong Ding, Chengjie Wang, Feiyue Huang, Dongyang Li, Wenshuai Qi, and Rongrong Ji. Aurora guard: Real-time face anti-spoofing via light reflection. *CoRR*, abs/1902.10311, 2019. 2

[21] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with restarts. *CoRR*, abs/1608.03983, 2016. 4

[22] Jukka Määttä, Abdenour Hadid, and Matti Pietikäinen. Face spoofing detection from single images using micro-texture analysis. In *Proc. IJCB*, pages 1–7, 2011. 1, 2

[23] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Synrhythm: Learning a deep heart rate estimator from general to specific. In *Proc. ICPR*, pages 3580–3585, 2018. 2

[24] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video. *arXiv preprint arXiv:1810.04927*, 2018. 2

[25] Gang Pan, Lin Sun, Zhaohui Wu, and Shihong Lao. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In *Proc. ICCV*, 2007. 2

[26] Keyurkumar Patel, Hu Han, and Anil K Jain. Cross-database face antispoofing with robust feature representation. In *Proc. CCBR*, pages 611–619, 2016. 2

[27] Keyurkumar Patel, Hu Han, and Anil K Jain. Secure face unlock: Spoof detection on smartphones. *IEEE Trans. Inf. Forensics Security*, 11(10):2268–2283, 2016. 1, 2

[28] Keyurkumar Patel, Hu Han, Anil K Jain, and Greg Ott. Live face video vs. spoof face video: Use of moiré patterns to detect replay video attacks. In *Proc. ICB*, pages 98–105, 2015. 1

[29] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proc. CVPR*, pages 234–778, 2004. 1

[30] Zezheng Wang, Chenxu Zhao, Yunxiao Qin, Qiusheng Zhou, and Zhen Lei. Exploiting temporal and depth information for multi-frame face anti-spoofing. *CoRR*, abs/1811.05118, 2018. 2

[31] Di Wen, Hu Han, and Anil K Jain. Face spoof detection with image distortion analysis. *IEEE Trans. Inf. Forensics Security*, 10(4):746–761, 2015. 1, 2, 4

[32] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Proc. ECCV*, pages 499–515, 2016. 2

[33] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proc. ECCV*, pages 3–19, 2018. 3

[34] Zhenqi Xu, Shan Li, and Weihong Deng. Learning temporal features using LSTM-CNN architecture for face anti-spoofing. In *Proc. ACPR*, pages 141–145, 2015. 2

[35] Jianwei Yang, Zhen Lei, and Stan Z Li. Learn convolutional neural network for face anti-spoofing. *arXiv preprint, arXiv:1408.5601*, 2014. 2

[36] Shifeng Zhang, Xiaobo Wang, Ajian Liu, Chenxu Zhao, Jun Wan, Sergio Escalera, Hailin Shi, Zezheng Wang, and Stan Z. Li. CASIA-SURF: A dataset and benchmark for large-scale multi-modal face anti-spoofing. In *Proc. CVPR*, 2019. 2, 3, 4, 5

[37] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z Li. A face antispoofing database with diverse attacks. In *Proc. ICB*, pages 26–31, 2012. 2, 4