2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops

Recognizing Multi-Modal Face Spoofing with Face Recognition Networks

Aleksandr Parkin
VisionLabs
a.parkin@visionlabs.ai

Oleg Grinchuk
VisionLabs
o.grinchuk@visionlabs.ai

Abstract

Detecting spoofing attacks plays a vital role for deploying automatic face recognition for biometric authentication in applications such as access control, face payment, device unlock, etc. In this paper we propose a new anti-spoofing network architecture that takes advantage of multi-modal image data and aggregates intra-channel features at multiple network layers. We also transfer strong facial features learned for face recognition and show their benefits for detecting spoofing attacks. Finally, to increase the generalization ability of our method to unseen attacks, we use an ensemble of models trained separately for distinct types of spoofing attacks. The proposed method achieves state-of-the-art result on the largest multi-modal anti-spoofing dataset CASIA-SURF [26].

1. Introduction

With the rapid growth of face recognition technology, it becomes crucial to protect automatic authentication systems from spoofing attacks and to deny unauthorized access. The deployed systems should be able to determine the liveness of the person in front of the camera, for example, by recognizing and denying any types of face presentation attacks such as printed photographs, video replays, 3D masks and others.

Face recognition has achieved tremendous progress with state-of-the-art methods readily by-passing human-level performance [22]. A significant part of this success can be attributed to the availability of large annotated face datasets [9, 19] typically collected from the Internet. On the contrary, datasets for face anti-spoofing attacks require a tedious process of manual data collection, and are therefore limited in the number of unique people and samples. Anti-spoofing algorithms, however, can benefit from different image modalities such as Infrared and Depth channels, captured by dedicated cameras. These modalities provide complementary information, hence, their combination is expected to improve liveness detection. Indeed, IR cameras are insensitive to electronic displays and can prevent attacks from phones and tablets, while depth channel makes it easier to distinguish flat printed surfaces from face shapes.

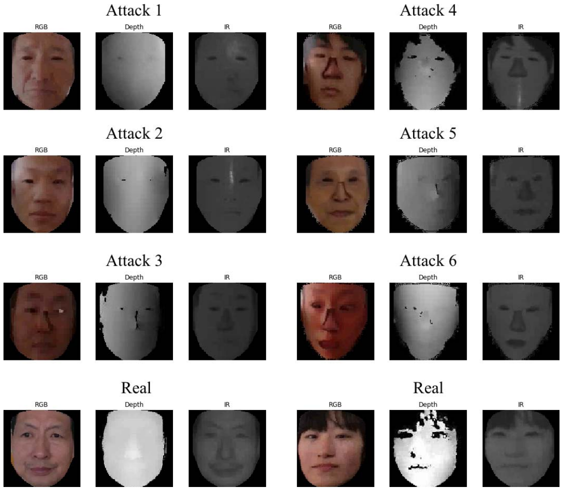

Figure 1. Examples of real and fake images from CASIA-SURF dataset.

The majority of current face anti-spoofing datasets contain RGB images only [2, 5]. Previous multi-modal datasets are limited in the number of subjects [4, 7], hence, the existing methods risk to overfit to the training data. The recently released face anti-spoofing dataset CASIA-SURF [26] pushes the limits of the liveness detection task both in terms of the dataset size and the number of presented modalities (RGB, IR, Depth). CASIA-SURF allows to develop new neural network models that benefit from multiple modalities and can be trained from a large corpus of data.

In this paper we introduce a new method for solving face anti-spoofing problem. We modify the architecture presented in [26], processing each modality separately and aggregating layer features at different levels, which increases the cooperation between RGB, IR and Depth branches of the neural network.

978-1-7281-2506-0/19/$31.00 ©2019 IEEE
DOI 10.1109/CVPRW.2019.00204

Despite being significantly larger than previous anti-spoofing datasets, CASIA-SURF is still orders of magnitude smaller compared to standard datasets for face recognition. To deploy powerful Convolutional Neural Network (CNN) models, we benefit from CNNs pretrained on four face attribute/identity recognition datasets and then fine-tune our final model on CASIA-SURF. We argue that such pre-training on different source domains provides rich face-specific features and can improve models for face anti-spoofing.

To increase the robustness to unknown attacks, we train multiple models on different subsets of the training data, leaving one attack type out, and then ensemble these models. As a result, our method achieves 99.8739 TPR at FPR=10^{-4} on the CASIA-SURF test set, ranking 1st in the Chalearn LAP multi-modal face anti-spoofing attack detection challenge [14].

2. Related work

There are two types of approaches to liveness detection for biometric security systems. The first family of methods [18, 23, 1] requires interaction with the user in the form of certain actions, such as eye blinking, head movements or changing facial expressions. While such methods may perform well to prevent spoofing attacks with printed images and rigid 3D masks, they can be compromised by video attacks demonstrating required actions. Moreover, such methods can be inappropriate or undesired in certain application scenarios.

The second family of methods [8, 21, 13, 15] is aimed at detecting liveness from just a single image of a person. Such algorithms are convenient and fast for end-users while the liveness verification could be run in the background.

There is a number of datasets with fake and real images that could be used for developing non-cooperative liveness detection. Replay-Attack [3], CASIA-FASD [27] and SiW [15] datasets contain still RGB images. MSU-MFSD [24], Replay-Mobile [5] and OULU-NPU [2] provide video recordings of attacks from mobile devices. However, these datasets contain only the RGB modality, hence, limiting the power and generalization of associated methods. As new types of spoofing attacks emerge (3D realistic silicon masks) and the quality of video devices is constantly improving, RGB channel is not efficient enough to provide a high level of security. Additional image channels, provided by special cameras, such as infrared or depth, could enrich the number of useful features (light distribution, eye reflection, face surface) therefore making anti-spoofing models more reliable.

	Surface	Eyes	Nose	Mouth	Split
1	Flat	✓			val/test
2	Curved	✓			val/test
3	Flat	✓	✓		val/test
4	Curved	✓	✓		train
5	Flat	✓	✓	✓	train
6	Curved	✓	✓	✓	train

Table 1. Different types of spoofing attacks from CASIA-SURF. All attacks contain printed images of the target, the paper prints could be bent or left flat while some regions could be cut out. Checkmarks indicate which regions were cut in each attack.

3. CASIA-SURF dataset

CASIA-SURF dataset [26] includes 21000 videos of 1000 subjects with one real and six fake videos per subject where each fake videos belongs to a different type of attack. The videos are recorded by the Intel RealSense SR300 camera and contain three modalities (RGB, IR and Depth). The dataset is divided into training, development and test subsets each containing 300, 100 and 600 unique subjects. Every 10th frame from each video is selected and distributed over the three subsets with 148K, 48K and 295K frames respectively.

To focus on the generalization to unknown attacks, Chalearn LAP challenge provides only a part of the CASIA-SURF dataset for training, i.e. for each subject only a subset of fake frames is available. Hence, challenge participants were given with about 30K frames for training and 9.6K frames for validation.

Examples of fake and real images from the CASIA-SURF dataset are illustrated in Fig. 1. Printed attacks differ by the shape (flat or curved) and regions showing parts of the real face. More information on the type of attacks is given in Table 1. Note, that attacks in the test set differ from attacks in the training set, therefore the successful model should avoid intra-class overfitting which was a common issue in previous anti-spoofing datasets.

3.1. Baseline method

With the release of the CASIA-SURF dataset, Zhang et al. [26] also introduced a method for the multi-modal face anti-spoofing task. The proposed pipeline is processing each of the three modalities separately using resnet-18 [10] as a backbone, and then performs the re-weighting of features from the last layer of each branch to select the more informative channel features while suppressing the less useful ones. Then the re-weighted features are concatenated and processed by two more residual blocks. Finally, the global average pooling (GAP) and two consecutive fully-connected layers complete the network structure. Authors provide extensive experiments to show the advantages of

Figure 2. The proposed architecture. RGB, IR and Depth streams are processed separately using res1, res2, res3 blocks from resnet-34 as a backbone. The res3 output features are re-weighted and fused via the squeeze and excitation (SE) block and then fed into res4. In addition, branch features from res1, res2, res3 are concatenated and processed by corresponding aggregation blocks, each aggregation block also uses information from the previous one. The resulting features from agg3 are fed into res4 and summed up with the features from the modality branch. On the diagram: GAP - global average pooling; $\bigoplus$ - concatenation; $+$ - elementwise addition.

this architecture and in our work we refer to it as a baseline method.

3.2. Evaluation metrics

The evaluation of face anti-spoofing methods can be done in different ways. Prior work [2, 5, 16, 11] uses Average Classification Error Rate (ACER) since liveness prediction can be seen as a binary classification task. However, similar to face recognition, one should pay attention to the True Positive Rate at some fixed False Positive Rate. This approach enables to measure how many real samples will pass the anti-spoofing test while accepting no more than some percentage of spoofing attacks.

Here we follow the evaluation metric of the Chalearn LAP Challenge, and report TPR at 10^{-4} FPR, which can be obtained from the receiver operating characteristic (ROC) on the target set.

4. Proposed method

In this section we describe our method and its training details.

4.1. Attack specific folds

To increase robustness to new attacks, where attack types at test time can differ from attacks presented in the training set, we split training data into three folds. Each fold contains two different attacks, while images of the third attack type are used for validation. Once trained, we treat three different networks as a single model by simply averaging their prediction scores.

	Backbone	Dataset	Task
1	resnet-34	Casia-Web face [25]	Face rec.
2	resnet-34	AFAD-lite [17]	Gender class.
3	resnet-50	MSCeleb-1M [9]	Face rec.
4	resnet-50	Asian dataset [28]	Face rec.

Table 2. Face datasets and CNN architectures used to pre-train our networks.

4.2. Transfer learning

Many image recognition tasks with limited training data benefit from CNN pre-training on large-scale image datasets, such as ImageNet [6]. Finetuning network parameters that have been pre-trained on various source tasks leads to different results on the target task. In our experiments we use four datasets designed for face recognition and gender classification (see Table 2), to create good initialization for our face anti-spoofing networks. We also use multiple backbone ResNet architectures and losses for initial tasks to increase the variability. Similar to networks trained for attack-specific folds in Sec. 4.1, we average predictions of models trained with different initialization.

Method	Initialization	Fold	TPR at FPR=10^{-4}
Zhang, Wang et al.[26]			56.80*
resnet-18		subject 5-fold	60.54
resnet-34		subject 5-fold	74.55
resnet-34		attack 3-fold	78.89
resnet-34	ImageNet [6]	attack 3-fold	92.12
resnet-34	CASIA-Web face [25]	attack 3-fold	99.80
A. resnet-34 with MLFA	CASIA-Web face [25]	attack 3-fold	99.87
B. resnet-50 with MLFA	MSCeleb-1M [9]	attack 3-fold	99.63
C. resnet-50 with MLFA	Asian dataset [28]	attack 3-fold	99.33
D. resnet-34 with MLFA	AFAD-lite [17]	attack 3-fold	98.70
A,B,C,D ensemble		attack 3-fold	100.00

Table 3. Results on CASIA-SURF validation subset.

4.3. Model architecture

Our final network architecture is based on the ResNet-34 and ResNet-50 [10] backbone with SE modules as illustrated in Fig. 2. Following the method described in [26], each modality is processed by the first three residual convolutional blocks, then the output features are fused using squeeze and excitation fusion module and processed by the remaining residual block. Differently from the baseline method we enrich the model with additional aggregation blocks at each feature level. Each aggregation block takes features from the corresponding residual blocks and from previous aggregation block, making model capable of finding inter-modal correlations not only at a fine level but also at a coarse one.

Additionally, we train each model using two initial random seeds. Given separate networks for attack-specific folds and different pre-trained models, our final liveness score is obtained by averaging outputs of 24 neural network.

5. Experiments

This section describes the implementation, hardware and software details and shows the importance of each additional improvement in terms of evaluation metric on the validation set, provided by Chalearn LAP challenge. We release the code and our trained models at https://github.com/AlexanderParkin/ChaLearn_liveness_challenge.

5.1. Implementation details

All the code was implemented in PyTorch [20] and models were trained on 4 NVIDIA 1080Ti. Single model trains about 3 hours and the inference takes 8 seconds per 1000 images.

All neural nets were trained using ADAM [12] with cosine learning rate strategy and optimized for standard cross entropy loss for two classes. We trained each model for 30 epochs with initial learning rate at 0.1 with batch size of 128. The same learning strategy was applied to pre-train models on the gender recognition task.

5.2. Preprocessing

CASIA-SURF already provides face crops so no detection algorithms were used to align images. Face crops were resized to 125 × 125 pixels and then center crop 112 × 112 was taken. At the training stage horizontal flip was applied with 0.5 probability. We also tested different crop and rotation strategies as well as test-time augmentation, however, this did not result in significant improvements and no additional augmentation was used in the final model except the above.

5.3. Baseline

Unless mentioned explicitly, we report results on Chalearn LAP challange validation set obtained from the Codalab evaluation platform. First of all, we reproduced baseline method [26] with Resnet-18 backbone and trained it using 5 fold cross-validation strategy. All folds are split based on the subject identity so images from the same person belong only to one fold. Then the score is averaged for the five trained nets and $TPR@FPR = 10^{-4}$ is reported in Table 3. The resulting performance is close to perfect and similar to the previously reported results in [26], which was calculated on the test set. The test set differs from the validation, but belongs to the same spoofing attack distribution.

Next, we expand the backbone architecture to ResNet-34 which improves the score by a large margin. Due to the GPU limitations we further focuse only on ResNet-34 and add Resnet-50 only at the final stage.

5.4. Attack-specific folds

Here we compare the 5-fold split strategy based on subject ids with the strategy based on spoof attack types. Real

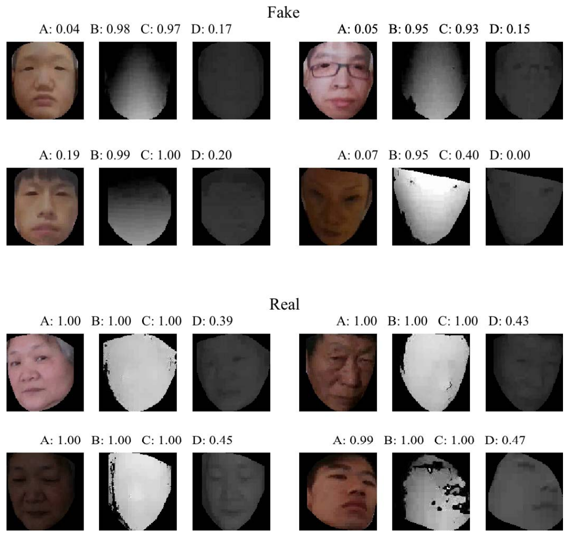
Fake
A: 0.04 B: 0.98 C: 0.97 D: 0.17
A: 0.05 B: 0.95 C: 0.93 D: 0.15
A: 0.19 B: 0.99 C: 1.00 D: 0.20
A: 0.07 B: 0.95 C: 0.40 D: 0.00
Real
A: 1.00 B: 1.00 C: 1.00 D: 0.39
A: 1.00 B: 1.00 C: 1.00 D: 0.43
A: 1.00 B: 1.00 C: 1.00 D: 0.45
A: 0.99 B: 1.00 C: 1.00 D: 0.47

Figure 3. Examples of fake and real samples with highest standard deviation among predicted liveness scores from models A,B,C,D.

examples by subject identity were assigned randomly to the one of the three folds.

Despite the fact that the new model computes an average of three network outputs while each of these networks was trained on less data compared to the subject 5-fold learning strategy, our model achieves better performance compared to baselines (see Table 3). We explain this by the improved generalization to new attacks due to the training for different types of attacks.

5.5. Initialization matters

In the next experiment we initialize each of the three modality branches of our network with the res1, res2, res3 blocks from the ImageNet pre-trained network [6]. The Fusion SE parts are left unchanged and the final res4 block is also initialized by the ImageNet pre-trained weights. Finetuning of this model on the CASIA-SURF dataset gives significant improvement over networks with random initialization (see Table 3). Moreover, switching pre-training to the face recognition task on the CASIA-Web face dataset [25] improves results by even a larger margin and reaches almost perfect TPR of 99.80%.

5.6. Multi-level feature aggregation

Here we examine the effect of multi-level feature aggregation (MLFA) described in the model architecture section. We initialize aggregation modules with random weights and train the new architecture following our best learning protocol. Our ResNet-34 network with MLFA blocks has demonstrated error reduction by the factor 1.5x compared to the network without MFLA blocks.

5.7. Ensembling

To improve the stability of our solution we use four face related datasets as an inialization for the final model. We used publicly available networks with weights trained for face recognition tasks on the CASIA-WebFace [25], MSCeleb-1M [9] and private asian faces [28]. We also trained a network for gender classification on the AFAD-lite [17] dataset. Different tasks, losses and datasets imply different convolutional features and the average prediction of models finetuned with such initializations leads to 100.00% TPR@FPR=10^{-4}.

Such a high score meets the requirements of real world security applications, however, it was achieved using a large number of ensembling networks. In future work we plan to focus on reducing the size of the model and making it applicable for the real-time execution.

5.8. Solution stability

The consistency and stability of model performance on unseen data is important especially when it comes to real world security applications. During the validation phase of the challenge seven teams achieved perfect or near perfect accuracy, however only three solutions managed to hold close level of performance on the test set (see Table 4), where ours showed the smallest drop in performance compared to the validation results.

	Valid	Test
Ours	100.0000	99.8739
Team 2	100.0000	99.8282
Team 3	100.0000	99.8052
Team 4	100.0000	98.1441
Team 5	99.9665	93.1550
Team 6	100.0000	87.2094
Team 7	100.0000	25.0601

Table 4. Shrinkage of TPR at FPR=10^{-4} score on validation and test sets of Chalearn LAP face anti-spoofing challenge.

We believe that the stability of our solution was caused by the diversity of networks in our final ensemble in terms of network architectures, pre-training tasks and random seeds.

5.9. Qualitative results

In this section we analyze difficult examples for our model. We run four networks (namely A,B,C,D in Table 3) on the Chalearn LAP challenge validation set and select examples with highest standard deviation on the liveness score among all samples. High STD implies conflicting predictions by different models. Fig. 3 shows examples for which the networks disagree at most. As can be seen, the model

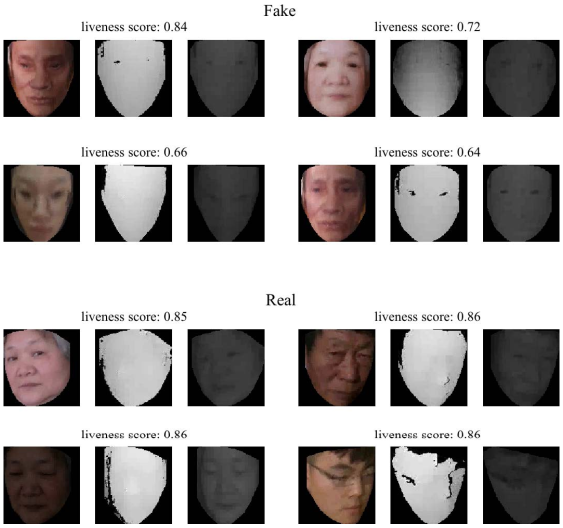

Figure 4. Examples of fake and real samples from validation subset where predicted liveness score is close to the threshold at FPR=10^{-4}.

D (which achieves the lowest TPR among all four models) tends to understate the liveness score, assigning reals to fakes. But it is helpful in the case of hard fake examples, when two out of three other networks are wrong. Therefore, Using only three models in the final ensemble would have led to lower score on the validation set.

Fig. 4 demonstrates fakes and real samples which were close to the threshold at FPR=10^{-4}. While they are distinguishable by human eye, one of the three modalities for every example looks similar to the normal one from the opposed class, so models based only on one modality may produce wrong predictions. Processing RGB, Depth and IR channels together allows to overcome this issue.

5.10. Multi-modality

Finally, we examine the advantage of multi-modal networks over networks trained for each of the three modalities separately. We take our architecture with three branches and aggregation blocks, but instead of passing (RGB, IR, Depth) channels, we trained three models with (RGB, RGB, RGB), (IR, IR, IR) and (Depth, Depth, Depth) inputs. This allows a fair comparison with multi-modal network since all these architectures were identical and had the same number of parameters.

As can be seen from Table 5, using only RGB images results in low performance. The corresponding model overfitted to the training set and achieved only 7.85% TPR at FPR=10^{-4}. The IR based model showed remarkably better results, reaching 57.41% TPR at FPR=10^{-4} since IR images contained less identity details and the dataset size in this case was not so crucial as it was for the RGB model. The highest score of 98.40% TPR at FPR=10^{-4} was achieved by the Depth modality, suggesting the importance of the facial shape information for the anti-spoofing task.

Modality	TPR at FPR		
	$= 10^{-2}$	$= 10^{-3}$	$= 10^{-4}$
RGB	71.74	22.34	7.85
IR	91.82	72.25	57.41
Depth	100.00	99.77	98.40
RGB+IR+Depth	100.00	100.00	99.87

Table 5. The effect of modalities measured on the validation set. All models were pre-trained on the CASIA-Web face recognition task and finetuned with the same learning protocol.

However, the multi-modal network performed much better than the Depth network alone, reducing false rejection error from 1.6% to 0.13%, and showing the evidence of the synergetic effect of modality fusion.

6. Conclusion

In this paper we have presented a new method for face anti-spoofing detection which has achieved top-1 rank at the Chalearn LAP face anti-spoofing challenge. We discussed in details three different directions: data, architecture and initialization, that summed up to a consistent solution, demonstrating significant improvements on a test set. First, we have demonstrated that careful selection of a training subset by the types of spoofing samples better generalizes to unseen attacks. Second, we have proposed a multi-level feature aggregation module which fully utilizes the feature fusion from different modalities both at coarse and fine levels. Finally, we have examined the influence of feature transfer from different pre-trained models on the target task and showed that using the ensemble of various face related tasks as source domains increases the stability and the performance of the system.

The code and pre-trained models for our approach are publicly available from the github repository at https://github.com/AlexanderParkin/ChaLearn_liveness_challenge.

References

[1] Samarth Bharadwaj, Tejas I Dhamecha, Mayank Vatsa, and Richa Singh. Face live detection method based on physiological motion analysis. *CVPR*, 2013. 2

[2] Zinelabinde Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdenour Hadid. Oulu-npu: A mobile face presentation attack database with real-world variations. *FG*, 2017. 1, 2, 3

[3] Ivana Chingovska, Andre Anjos, and Sebastien Marcel. On the effectiveness of local binary patterns in face antispoofing. *BIOSIG*, 2012. 2

[4] Ivana Chingovska, Nesli Erdogmus, Andre Anjos, and Sebastien Marcel. Face recognition systems under spoofing attacks. *Face Recognition Across the Imaging Spectrum*, 2016. 1

[5] Artur Costa-Pazo, Sushil Bhattacharjee, Esteban Vazquez-Fernandez, and Sebastien Marcel. The replay-mobile face presentation-attack database. *BIOSIG*, 2016. 1, 2, 3

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 3, 4, 5

[7] Nesli Erdogmus and Sebastien Marcel. Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect. *BTAS*, 2014. 1

[8] Litong Feng, Lai-Man Po, Yuming Li, Xuyuan Xu, Fang Yuan, Terence Chun-Ho Cheung, and Kwok-Wai Cheung. Integration of image quality and motion cues for face anti-spoofing: A neural network approach. *JVCIR*, 2016. 2

[9] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-Celeb-1M: A dataset and benchmark for large scale face recognition. In *European Conference on Computer Vision*. Springer, 2016. 1, 3, 4, 5

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2, 4

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

[12] Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 4

[13] Lei Li, Xiaoyi Feng, Zinelabidine Boulkenafet, Zhaoqiang Xia, Mingming Li, and Abdenour Hadid. An original face anti-spoofing approach using partial convolutional neural network. *IPTA*, 2016. 2

[14] 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. *CVPR workshop*, 2019. 2

[15] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. *CVPR*, 2018. 2

[16] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. *CVPR*, 2018. 3

[17] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. *CVPR*, 2016. 3, 4, 5

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

[19] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015. 1

[20] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 4

[21] Keyurkumar Patel, Hu Han, and Anil K Jain. Secure face unlock: Spoof detection on smartphones. *TIFS*, 2016. 2

[22] P Jonathon Phillips, Amy N Yates, Ying Hu, Carina A Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G Cavazos, Géraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, et al. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24):6171–6176, 2018. 1

[23] Liting Wang, Xiaoqing Ding, and Chi Fang. Face live detection method based on physiological motion analysis. *Tsinghua Science and Technology*, 2009. 2

[24] Di Wen, Hu Han, and Anil K Jain. Face spoof detection with image distortion analysis. *TIFS*, 2015. 2

[25] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Learning face representation from scratch. *arXiv*, 2014. 3, 4, 5

[26] Shifeng Zhang, Xiaobo Wang, Ajian Liu, Chenxu Zhao, Jun Wan, Sergio Escalera, Hailin Shi, Zezheng Wang, and Stan Z. Li. A dataset and benchmark for large-scale multi-modal face anti-spoofing. *CVPR*, 2019. 1, 2, 4

[27] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z Li. A face antispoofing database with diverse attacks. *ICB*, 2012. 2

[28] Jian Zhao, Yu Cheng, Yan Xu, Lin Xiong, Jianshu Li, Fang Zhao, Karlekar Jayashree, Sugiri Pranata, Shengmei Shen, Junliang Xing, et al. Towards pose invariant face recognition in the wild. In *CVPR*, pages 2207–2216, 2018. 3, 4, 5