FaceBagNet: Bag-of-local-features Model for Multi-modal Face Anti-spoofing

Tao Shen
ReadSense
Shanghai, China
taoshen@readsense.cn

Yuyu Huang
ReadSense
Shanghai, China
huangyuyu@stu.xmu.edu.cn

Zhijun Tong
ReadSense
Shanghai, China
thomas@readsense.cn

Abstract

Face anti-spoofing detection is a crucial procedure in biometric face recognition systems. State-of-the-art approaches, based on Convolutional Neural Networks (CNNs), present good results in this field. However, previous works focus on one single modal data with limited number of subjects. The recently published CASIA-SURF dataset is the largest dataset that consists of 1000 subjects and 21000 video clips with 3 modalities (RGB, Depth and IR). In this paper, we propose a multi-stream CNN architecture called FaceBagNet to make full use of this data. The input of FaceBagNet is patch-level images which contributes to extract spoof-specific discriminative information. In addition, in order to prevent overfitting and for better learning the fusion features, we design a Modal Feature Erasing (MFE) operation on the multi-modal features which erases features from one randomly selected modality during training. As the result, our approach wins the second place in CVPR 2019 ChaLearn Face Anti-spoofing attack detection challenge. Our final submission gets the score of 99.8052% (TPR@FPR = 10e-4) on the test set.

1. Introduction

The face image is the most accessible biometric modality which is used for highly accurate face recognition systems, while it is vulnerable to many different types of presentation attacks[7, 1]. Consequently, face anti-spoofing is an important task in the area of computer vision and attracts lots of attentions for its application in face recognition systems[20, 24, 6]. It aims to determine whether the captured face is real.

For traditional methods, most researchers utilized handcrafted features, such as LBP[19, 29, 6, 17], HoG [29], SIFT[21], SURF and DoG[13] to learn different feature distributions between live faces and spoof ones. Boulkenafet et al.[3] used LBP features to characterize color-texture. These color-texture feature-vectors can be classified by using support vector machines (SVM).

Recently, CNN-based methods[9, 22] are presented in face presentation attack detection(PAD) community. They treat face PAD as a binary classification problem. Some researchers apply CNN as a feature extractor[21] and extract discriminative features to distinguish between live and spoofing. Yang et al.[28] proposed a CNN with the same architecture as ImageNet. Atoum et al.[18] compared the performances of three different CNN architectures: the Inception-v3 [23] and two versions of ResNet [10], namely ResNet50 (a 50-layer ResNet) and ResNet152 (the 152-layer version). [2] utilized an ensemble of patch-based and depth-based CNN in facial unlocking systems. Feng et al.[9] proposed to use multiple cues as the CNN input for live/spoof classification. In [25], Tu et al. proposed an LSTM-CNN architecture to conduct a joint prediction for multiple frames of a video.

All these methods proved that the CNNs[28, 16, 12] can be used very effectively for the face anti-spoofing by automatically extracting the useful features from training data. Unfortunately, the existing face anti-spoofing datasets have limited number of subjects[31, 6, 4], which greatly decreased the generalization ability of these methods. Previous published works are mainly based on these datasets which are hardly to meet the requirements of practical applications. The recently published CASIA-SURF[30, 15] dataset consists of 1000 subjects and 21000 video clips with 3 modalities (RGB, Depth and IR). It is the largest face anti-spoofing dataset in term of the number of subjects and videos. With the help of the CASIA-SURF[30, 15] dataset, we design a multi-stream network to classify multi-modal face images based on image patches. Our approach is motivated by BagNets[5], which classifies images by using small local image patches. Although this strategy ignores some spatial information, it reaches a surprisingly high accuracy on ImageNet. The main works and contributions of this paper are as follows:

(1) A patch-based features learning method. It classifies a face image based on the occurrences of small local image features which shows strong performance.

(2) A multi-stream fusion method with our Modal Fea-

ture Erasing (MFE) which integrates the diverse information involved in sub-features. It improves the performances considerably and demonstrates itself as an effective method for muti-modal face anti-spoofing.

2. Related Work

The existing face anti-spoofing methods generally can be categorized into two groups: (1) traditional face anti-spoofing methods, (2) CNN-based face anti-spoofing methods.

Traditional face anti-spoofing methods: Many prior works utilize hand-crafted features, such as LBP, HoG, SIFT and SURF, and usually adopt traditional classifiers such as SVM and LDA. In order to overcome the influence of illumination variations, Zhang et al.[31] utilized the multiple DoG filters to remove the noise and low-frequency information. They used SVM classifier to distinguish the genuine and fake faces. Pereira et at.[8] used the space and time descriptors to encode rich information. Chingovska et al.[6] used LBP descriptors to extract discriminative features from a greyscale image and then applied 3 classifiers to perform it as the classification problem. Since the traditional methods are sensitive to different illuminations, poses and specific identities, these methods could not capture discriminative representations and their generalization abilities are poor.

CNN-based face anti-spoofing methods: Recently in many visual information processing areas like object detection, image classification, image captioning and semantic segmentation, CNN has been proven to be an effective method. Therefore, CNN is widely used in face anti-spoofing and liveness detection. Atoum et al.[27] proposed two-stream CNN based face anti-spoofing methods including patch-based and depth-based. Li et al.[14] extracted the features and applied the principle component analysis to improve the robustness of the face recognition system. In [9], Feng et al. designed different face images to feed them into CNN, and then directly classified whether the face is real. In [14], Li et al. finetuned the CNN over the face anti-spoofing datasets and achieved high performance. In [27], Xu et al. firstly introduced the LSTMs in face anti-spoofing area, they both used the local features and temporal features which can be learned and sorted in LSTM units. Liu et al.[16] designed a novel network architecture to leverage the Depth map and rPPG signal as supervision with the goal of improving generalization capability. Amin et al.[12] introduced a new perspective for solving the face anti-spoofing by inversely decomposing a spoof face into the live face and the spoof noise pattern.

Overall, prior methods regard face anti-spoofing as a binary classification problem, and they cannot generalize well due to the over-fitting to training data. In [5], Brendel et al. extracted patch features from the input image and then achieved remarkable improvements over datasets. This work brings us the insight that we need to involve the local features to solve face anti-spoofing problem.

Table 1. Architecture of the proposed FaceBagNet.

Patch size	Configuration	
layer1	conv 3×3, 32	
layer2	$conv1\times1$, 64 $conv3\times3$, 64, group 32, stride 2 $conv1\times1$, 128	$\times$ 2
layer3	$conv1\times1$, 128 $conv3\times3$, 128, group 32, stride 2 $conv1\times1$, 256	$\times$ 2
layer4	$conv1\times1$, 256 $conv3\times3$, 256, group 32, stride 2 $conv1\times1$, 512	$\times$ 2
layer5	$conv1\times1$, 512 $conv3\times3$, 512, group 32, stride 2 $conv1\times1$, 1024	$\times$ 2
layer6	global avg pooling fc, 2	

3. Methods

3.1. The overall architecture

In this work, we propose a multi-stream CNN architecture called FaceBagNet with Modal Feature Erasing (MFE) for multi-modal face anti-spoofing detection. Our method consists of two components, (1) patch-based features learning, (2) multi-stream fusion with MFE. For the patch-based features learning, we train a deep neural network by using patches randomly extracted from face images to learn rich appearance features. For the multi-stream fusion, features from different modalities are randomly erased during training, which are then fused to perform classification. Figure 1 shows a high-level illustration of three streams along with a fusion strategy for combining them.

3.2. Patch-based features learning

The spoof-specific discriminative information exists in the whole face area. Therefore, we can use the patch-level image to enforce CNN to extract such information. The usual patch-based approaches split the full face into several fixed non-overlapping regions. Then each patch is used to train an independent sub-network. In this paper, for each modality, we train one single CNN on random patches extracted from the faces. We use a self designed ResNext[26] network to extract deep features. The network consists of five group convolutional blocks, a global average pooling layer and a softmax layer. Table 1 presents the network architecture in terms of its layers, i.e., size of kernels, number of output feature maps, number of groups,

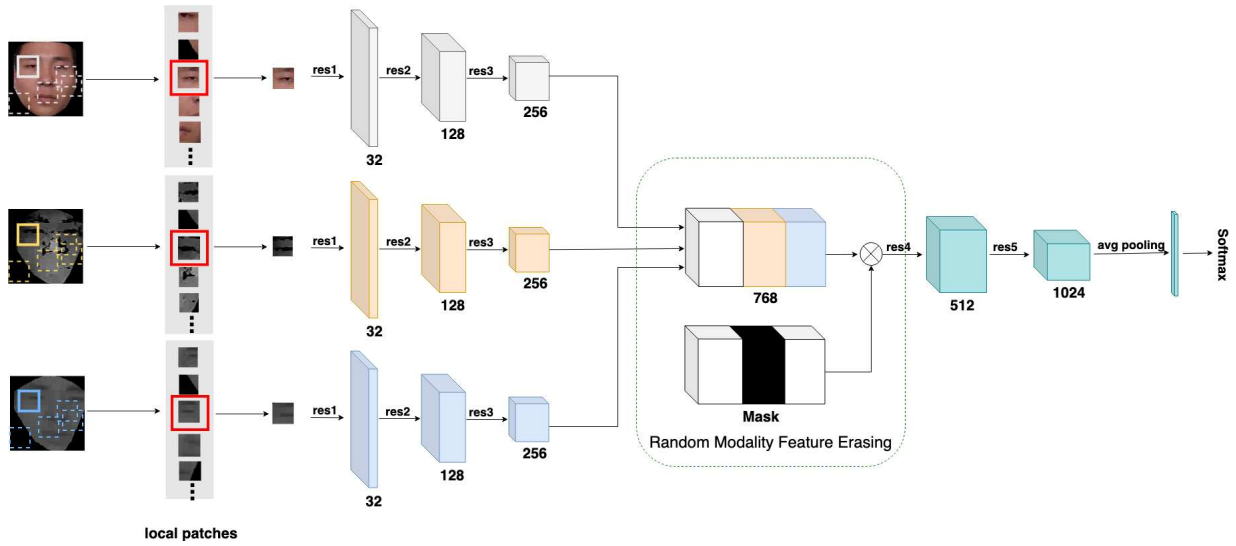

Figure 1. Architecture of the proposed face anti-spoofing approach. The fusion network is trained from scratch in which RGB, Depth and IR face patches are feed into it at the same time. Image augmentation is applied and modal features from sub-network are randomly erased during training.

strides. Our experiments show, the patch-based features are highly discriminative among different attacks. In the experiments section, quantitative results comparing different sizes of patches will be presented.

3.3. Multi-stream fusion with MFE

Since the feature distributions of different modalities are different, the proposed model makes efforts to exploit the interdependencies between different modalities as well. As shown in Figure 1, we use a multi-stream architecture with three sub-networks to perform multi-modal features fusion. We concatenate feature maps of three sub-networks after the third convolutional block (res3).

As studied in [30], directly concatenating features from each sub-network cannot make full use of the characteristics between different modalities. In order to prevent overfitting and for better learning the fusion features, we design a Modal Feature Erasing operation on the multi-modal features. For one batch of inputs, the concatenated feature tensor is computed by three sub networks. During training, the features from one randomly selected modal sub-network are erased and the corresponding units inside the erased area are set to zero. The fusion network is trained from scratch in which RGB, Depth and IR data are fed separately into each sub-network at the same time.

4. Experiments

4.1. Dataset and Evaluation Protocol

The CASIA-SURF[30, 15] dataset is currently the largest face anti-spoofing dataset including three modalities (i.e., RGB, Depth and IR), as shown in Figure 2. This

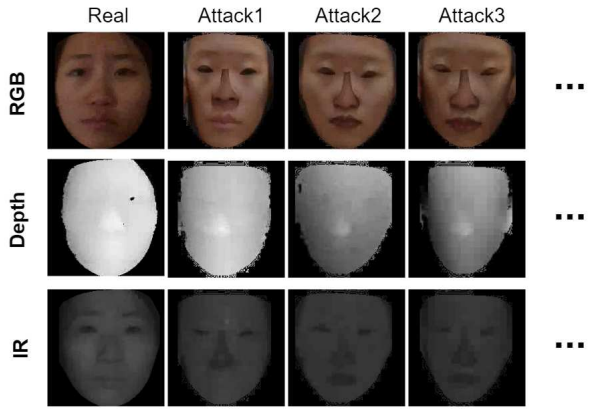

Figure 2. Examples of different attacks in CASIA-SURF dataset.

dataset contains 1000 Chinese people in 21000 videos and each sample includes 1 live video clip, and 6 fake video clips under different attack ways. The dataset is generated by 6 attacks. Eyes, nose, mouth areas or a combination of them are removed in different attack styles. The complex background is removed from the original videos except face areas. The dataset is separated into training set, validation set and test set. The training, validation and testing sets have 300, 100 and 600 subjects respectively. The complex backgrounds are removed from the original images except face areas. The three modality images are cropped and aligned. The resolution of RGB images in this dataset is 1280×720, and 640×480 for Depth, IR and aligned images. It is the most challenging dataset in face anti-spoofing area.

After training, Attack Presentation Classification Error Rate (APCER), Normal Presentation Classification Er-

Table 2. Test set results and rankings of the final stage teams in ChaLearn Face Anti-spoofing attack detection challenge, the best indicators are bold.

Team Name	FP	FN	APCER(%)	NPCER(%)	ACER(%)	TPR(%) @FPR=10e-2	TPR(%) @FPR=10e-3	TPR(%) @FPR=10e-4
VisionLabs	**3**	27	**0.0074**	0.1546	0.0810	99.9885	**99.9541**	**99.8739**
ReadSense(our team)	77	**1**	0.1912	**0.0057**	0.0985	**100.0000**	99.9472	99.8052
Feather	48	53	0.1192	0.1392	0.1292	99.9541	99.8396	98.1441
Hahahaha	55	214	0.1366	1.2257	0.6812	99.6849	98.5909	93.1550
MAC-adv-group	825	30	2.0495	0.1718	1.1107	99.5131	97.2505	89.5579

Table 3. The comparisons on different patch sizes and modalites. All models are trained in the CASIA-SURF training set and tested on the validation set.

Patch size	Modal	ACER	TPR@FPR = 10E-4
16×16	RGB	4.5	94.9
	Depth	2.0	98.0
	IR	1.9	96.2
	Fusion	1.5	98.4
32×32	RGB	4.2	95.8
	Depth	0.8	99.3
	IR	1.5	98.1
	Fusion	0.0	100.0
48×48	RGB	3.1	96.1
	Depth	0.2	99.8
	IR	1.2	98.6
	Fusion	0.1	99.9
96×96	RGB	13.8	81.2
	Depth	5.2	92.8
	IR	13.4	81.4
	Fusion	1.7	97.9
fullface	RGB	15.9	78.6
	Depth	8.8	88.6
	IR	11.3	84.3
	Fusion	4.8	93.7

ror Rate (NPCER) and Average Classification Error Rate (ACER) statistics are then calculated as evaluation results for our proposed model. According to the requirements of real applications, Receiver Operating Characteristic is chosen to select a threshold to trade off the false positive rate(FPR) and true positive rate(TPR).

4.2. Implementation details

The full face images are resized to size 112×112. We use random flipping, rotation, resizing, cropping for data augmentation. Patches are randomly extracted from the 112×112 full face images. All models are trained using one Titan X(Pascal) GPU with a batch size of 512. We use the Stochastic Gradient Descent (SGD) optimizer with a cyclic cosine annealing learning rate schedule[11]. The whole training procedure has 250 epochs and takes approximately 3 hours. Weight decay and momentum are set to 0.0005 and 0.9, respectively. We use PyTorch as the deep learning framework.

Table 4. The comparison on different training strategy. All models are trained with 32×32 size image patches.

Modal	ACER	TPR@FPR = 10E-4
Fusion(w.o CLR&MFE)	1.60	98.0
Fusion(w.o MFE)	0.60	98.5
Fusion(w.o CLR)	0.60	99.2
Fusion	0.00	100.0
Fusion(Erase RGB)	0.51	99.3
Fusion(Erase Depth)	0.49	99.4
Fusion(Erase IR)	0.84	99.3
Fusion	0.00	100.0

4.3. Results

To evaluate the effectiveness of our proposed model, we design several experiments with different configurations to make comparisons between them. The details of comparison experiments are presented as below:

The Effect of Patch Sizes and Modality: In this setting, we use different patch sizes in our model, i.e 16×16, 32×32, 48×48 and 64×64. For performance comparisons, all the models are inferred 36 times with 9 non-overlapping image patches and 4 flipped input. As depicted in Table 3, for single modal input, among the three modalities, the depth data achieves the best performance of 0.8% (ACER), TPR=99.3% @FPR=10e-4. Specifically, fusing all the three modalities has strong performance across all patch sizes. It can be concluded that our proposed method with fusion modality achieves the best results.

The Effect of Modal Feature Erasing and Training strategy: We investigate how the random modal feature erasing and training strategy affect model performance for face anti-spoofing. "w.o CLR" denotes that we use conventional SGD training with a standard decaying learning rate schedule until convergence instead of using cyclic learning rate. "w.o MFE" denotes that random modal features erasing are not applied. As shown in Table 4, both the cyclic learning rate and random modal feature erasing strategy are critical for achieving a high performance. After training the

fusion model, we erase features from one modal and then evaluate the performance. We evaluate the performance of the trained fusion model with single modal feature erasing. In table 4, from the validation score, we can conclude that the complementarity among different modalities can be learned to obtain better results.

Comparing with other teams in ChaLearn Face Anti-spoofing attack detection challenge: Our final submission in this challenge is an ensemble result which combined outputs of three models in different patch sizes (32×32, 48×48 and 64×64) and we ranked the second place in the end. We are the only team that did not use the full face image as model input. The result of FN = 1 shows that our patch based learning method can effectively prevent the model from misclassifying the real face into an attack one by comparing with other top ranked teams. As shown in Table 2, the results of the top three teams are significantly better than other teams on testing set. Especially, the TPR@FPR=10e-4 values of our team and VisionLabs are relatively close. Whereas, VisionLabs applied plentiful data from other tasks to pretrain the model, and our team only used a one-stage and end-to-end training schedule. Consequently, it also confirms the superiority of our solution.

5. Conclusions

In this paper, we propose a face anti-spoofing network based on Bag-of-local-features (named FaceBagNet) to determine whether the captured multi-modal face images are real. A patch-based feature learning method is used to extract discriminative information. Multi-stream fusion with MFE layer is applied to improve the performance. Our study demonstrates that both patch-based feature learning method and multi-stream fusion with MFE are effective methods for face anti-spoofing. Overall, our solution is simple but effective and easy to use in practical application scenarios. As the result, our approach wins the second place in CVPR 2019 ChaLearn Face Anti-spoofing attack detection challenge. Our final submission gets the score of 99.8052% (TPR@FPR = 10e-4) on the test set.

References

[1] A. Alotaibi and A. Mahmood. Deep face liveness detection based on nonlinear diffusion using convolution neural network. *Signal, Image and Video Processing*, 11(4):713–720, 2017.

[2] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu. Face anti-spoofing using patch and depth-based cnns. In *2017 IEEE International Joint Conference on Biometrics, IJCB 2017, Denver, CO, USA, October 1-4, 2017*, pages 319–328, 2017.

[3] P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors. *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, 2012.

[4] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid. OULU-NPU: A mobile face presentation attack database with real-world variations. In *12th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2017, Washington, DC, USA, May 30 - June 3, 2017*, pages 612–618, 2017.

[5] W. Brendel and M. Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *International Conference on Learning Representations*, 2019.

[6] I. Chingovska, A. Anjos, and S. Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *2012 BIOSIG - Proceedings of the International Conference of Biometrics Special Interest Group, Darmstadt, Germany, September 6-7, 2012*, pages 1–7, 2012.

[7] A. da Silva Pinto, W. R. Schwartz, H. Pedrini, and A. de Rezende Rocha. Using visual rhythms for detecting video-based facial spoof attacks. *IEEE Trans. Information Forensics and Security*, 10(5):1025–1038, 2015.

[8] T. de Freitas Pereira, A. Anjos, J. M. D. Martino, and S. Marcel. LBP - TOP based countermeasure against face spoofing attacks. In *Computer Vision - ACCV 2012 Workshops - ACCV 2012 International Workshops, Daejeon, Korea, November 5-6, 2012, Revised Selected Papers, Part I*, pages 121–132, 2012.

[9] L. Feng, L. Po, Y. Li, X. Xu, F. Yuan, T. C. Cheung, and K. Cheung. Integration of image quality and motion cues for face anti-spoofing: A neural network approach. *J. Visual Communication and Image Representation*, 38:451–460, 2016.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016.

[11] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger. Snapshot ensembles: Train 1, get M for free. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[12] A. Jourabloo, Y. Liu, and X. Liu. Face de-spoofing: Anti-spoofing via noise modeling. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII*, pages 297–315, 2018.

[13] W. Kim, S. Suh, and J. Han. Face liveness detection from a single image via diffusion speed model. *IEEE Trans. Image Processing*, 24(8):2456–2465, 2015.

[14] L. Li, X. Feng, Z. Boulkenafet, Z. Xia, M. Li, and A. Hadid. An original face anti-spoofing approach using partial convolutional neural network. In *Sixth International Conference on Image Processing Theory, Tools and Applications, IPTA 2016, Oulu, Finland, December 12-15, 2016*, pages 1–6, 2016.

[15] A. Liu, J. Wan, S. Escalera, H. J. Escalante, Z. Tan, Q. Yuan, K. Wang, G. Guo, I. Guyon, and S. Z. Li. Multi-modal face

anti-spoong attack detection challenge at cvpr2019. *CVPR workshop*, 2019.

[16] Y. Liu, A. Jourabloo, and X. Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. *CoRR*, abs/1803.11097, 2018.

[17] J. Määttä, A. Hadid, and M. Pietikäinen. Face spoofing detection from single images using texture and local shape analysis. *IET Biometrics*, 1(1):3–10, 2012.

[18] C. Nagpal and S. R. Dubey. A performance evaluation of convolutional neural networks for face anti spoofing. *CoRR*, abs/1805.04176, 2018.

[19] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7):971–987, 2002.

[20] G. Pan, L. Sun, Z. Wu, and S. Lao. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In *IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007*, pages 1–8, 2007.

[21] K. Patel, H. Han, and A. K. Jain. Secure face unlock: Spoof detection on smartphones. *IEEE Trans. Information Forensics and Security*, 11(10):2268–2283, 2016.

[22] K. Patel, H. Han, and A. K. Jain. Secure face unlock: Spoof detection on smartphones. *IEEE Trans. Information Forensics and Security*, 11(10):2268–2283, 2016.

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826, 2016.

[24] X. Tan, Y. Li, J. Liu, and L. Jiang. Face liveness detection from a single image with sparse low rank bilinear discriminative model. In *Computer Vision - ECCV 2010 - 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part VI*, pages 504–517, 2010.

[25] X. Tu, H. Zhang, M. Xie, Y. Luo, Y. Zhang, and Z. Ma. Enhance the motion cues for face anti-spoofing using CNN-LSTM architecture. *CoRR*, abs/1901.05635, 2019.

[26] S. Xie, R. B. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5987–5995, 2017.

[27] Z. Xu, S. Li, and W. Deng. Learning temporal features using LSTM-CNN architecture for face anti-spoofing. In *3rd IAPR Asian Conference on Pattern Recognition, ACPR 2015, Kuala Lumpur, Malaysia, November 3-6, 2015*, pages 141–145, 2015.

[28] J. Yang, Z. Lei, and S. Z. Li. Learn convolutional neural network for face anti-spoofing. *CoRR*, abs/1408.5601, 2014.

[29] J. Yang, Z. Lei, S. Liao, and S. Z. Li. Face liveness detection with component dependent descriptor. In *International Conference on Biometrics, ICB 2013, 4-7 June, 2013, Madrid, Spain*, pages 1–6, 2013.

[30] S. Zhang, X. Wang, A. Liu, C. Zhao, J. Wan, S. Escalera, H. Shi, Z. Wang, and S. Z. Li. CASIA-SURF: A dataset and benchmark for large-scale multi-modal face anti-spoofing. *CVPR*, 2019.

[31] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li. A face antispoofing database with diverse attacks. In *5th IAPR International Conference on Biometrics, ICB 2012, New Delhi, India, March 29 - April 1, 2012*, pages 26–31, 2012.