IEEE
THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS INC.
CAS
SIGNAL PROCESSING
ALGORITHMS, ARCHITECTURES, ARRANGEMENTS, AND APPLICATIONS
SPA 2017
September 20-22nd, 2017, Poznań, POLAND

Deep Learning for Face Anti-Spoofing: An End-to-End Approach

Yasar Abbas Ur Rehman*, Lai Man Po†, Mengyang Liu‡

Department of Electronic Engineering,
City University of Hong Kong,
Kowloon, Hong Kong
Email: {*yaurehman2-c, ‡mengyaliu7-c}@my.cityu.edu.hk, †eelmpo@cityu.edu.hk

***Abstract*—The importance of face anti-spoofing algorithms in biometric authentication systems is becoming indispensable. Recently, the success of Convolution Neural Networks (CNN) in key application areas of computer vision has encouraged its use in face biometrics for face anti-spoofing and verification applications. However, small training data has restricted the use of deep CNN architectures for face anti-spoofing applications. In this paper, we develop an end-to-end CNN architecture for face anti-spoofing application. i.e. a deep CNN architecture which directly map the raw input face images to the corresponding output classes. Additionally, an efficient training strategy has been proposed to enable the use of deeper CNN structures for face anti-spoofing applications and to enable the growth of training data in autonomous way. For training a CNN architecture, we propose a 50RS-30SeC-1E (50 Random Samples-30 Sub-epochs Count-1Epoch) training strategy. The training data is randomly sampled during each forward-pass through the CNN architecture and 30 such passes counts for 1 complete epoch. An 11-layer VGG network with 2 derived VGG-11 networks have been trained for face anti-spoofing on CASIA-FASD dataset. Experimental results show significant improvement on various face-spoofing scenarios. A 3% improvement over state of the art approaches has been reported for Overall Test (OT) while achieving a lowest EER of 5%.**

I. Introduction

Face anti-spoofing techniques are used for the recognition of live-face and fake-face. It has been one of the prevailing issue in todays biometric applications. Surveillance applications that rely on biometrics are based on (either solely or in combination of) face, fingerprint and iris recognition algorithms respectively. Therefore the robustness of biometrics algorithms to various types of known intrusion attacks or unknown intrusion attacks must be high. Apart from surveillance, face recognition algorithms have applications in banking and automatic transactions as well. Thus, face recognition algorithms developed for these applications must have a low tolerance for making errors [1]. However, face recognition systems are vulnerable to various types of face-spoofing attacks. The biometric communities have defined several types of face-spoofing attacks to aid the development of robust face anti-spoofing algorithms. For example, CASIA-FASD [2] database has provided training and testing set based on presentation attacks like wrapped-photo attacks, cut-photo attacks and video attacks. These and other types of presentation attacks to the face biometric systems [3] have enabled the biometric communities to propose various state of the art techniques for face anti-spoofing applications. Face anti-spoofing techniques can be broadly categorized into fixed feature based face anti-spoofing techniques [4] and learnable feature based face anti-spoofing techniques [5]. The fixed feature based face anti-spoofing techniques can be further categorized into texture based, motion based, 3D shape based and techniques based on multi-spectral reflectance. On the other hand, the learnable feature based face anti-spoofing techniques use Convolutional Neural Networks (CNN) that do not require fixed, hand crafted features. Instead, CNN based algorithms learn the features from the input data during training.

Face anti-spoofing is a specific type of face classification problem, where the input image to a system contains only face as an object as opposed to the image classification problem like ImageNet [6] that contain different objects in an image. To the best of our knowledge, there have been no work that use deep CNN architecture (an 11-layer network) using end-to-end learning for face anti-spoofing applications, i.e. a deep architecture which directly map the raw input face images to the corresponding output classes. In this paper, we develop an end-to-end architecture for face anti-spoofing application. An 11 layers VGG network (VGG-11) [7] and 2 networks derived from the VGG-11 network are trained from scratch on face anti-spoofing database. During training of a CNN network, at each forward-pass, a set of 50 face images are randomly sampled from whole video-data as an input to CNN network. Then, 30 such forward-pass through a CNN network counts for 1 complete epochs. We called this method as 50RS-30SeC-1E(50 Random Samples-30 Sub-epochs-Counts 1 Epoch). Thus, at the end of a single complete epoch we pass 1500 random face images (50 samples × 30 sub-epochs) through a CNN network. We utilize CASIA-FASD face anti-spoofing database for assessing the performance of the proposed CNN networks on the test protocol defined by [2]. We also study various techniques for performance comparison on the face anti-spoofing test dataset. Traditionally, for face anti-spoofing techniques, the performance of a system is evaluated by comparing the output probability of class against a set of threshold values between 0 and 1. We call this technique as threshold-operation here. On contrary, in CNN, the aim is to find the top-1 percent accuracy at the output of the network, i.e. Select the class with the highest probability and determine whether it matches with the corresponding ground-truth class label or not. Since at this point, keeping the spirit of both, the traditional approaches used for evaluating the performance of face anti-spoofing algorithms and the approaches proposed for evaluating CNN performance, we evaluate the performance of the proposed CNN networks, by using both the threshold-operation and the top-1 percent accuracy.

The rest of this paper has been organized as follows: Section II discusses about CNN networks and its use in face anti-spoofing algorithms. Section III present a detailed methodology for training a deep CNN network for face anti-spoofing application. Section IV reports the results obtained by training and testing the proposed CNN models on face anti-spoofing database. Finally, Section V provide some concluding remarks on the use of CNN for face anti-spoofing applications.

II. Literature Review

The remarkable success of Convolutional Neural Networks (CNN) [8] in ImageNet [6] competition has attracted a multitude of researchers in the computer vision community to investigate its potential latent capabilities in attaining such a high performance. The progressive improvement of CNN networks in general category of image classification [7], [8], [9], [10] and object detection[11], [12], [13], [14], [15] has opened the branches and potential application of CNN in other domains like face anti-spoofing.

Although, a vast amount of literature is available for fixed feature based face anti-spoofing algorithms such as [16], [17], [18], [19], very few algorithms were reported in the last few years that utilized CNN for face anti-spoofing applications[20], [21], [22], [23], [24]. Usually in these CNN based face anti-spoofing algorithms, the designer either used a pre-trained CNN network with Support Vector Machine (SVM) classifier or utilized a CNN network with a very low depth. For example, in [5], the authors proposed a shallow network for combined detection of face, iris and finger print spoofing attacks. Although their method gave satisfactory results, however a shallow network is unable to give abstract and high level features. A similar shallow network was proposed by authors in [21], [22]. In [23], the authors proposed a CNN network for face anti-spoofing to classify various attacks on two state of the art face anti-spoofing datasets, i.e. CASIA-FASD and REPLY-ATTACK datasets. In their method, a face region was first localized followed by data augmentation at five different scales before training a CNN network. However, after training CNN networks, the final features were used to train the Support Vector Machine (SVM) classifier for face anti-spoofing. As the method obtained remarkable results on CASIA-FASD and REPLY-ATTACK database, it cannot be regarded as an end-to-end learning. The approach in [24] exploited the layer-wise features of CNN to train an SVM algorithm. In their network, they used various shallow to deep VGG networks features for face anti-spoofing. However, they used a pre-trained network for face anti-spoofing and thus there are no details of training a CNN network. In this paper, an efficient end-to-end learning strategy is proposed to train a deep CNN network effectively for face anti-spoofing applications when the training data is limited.

III. Methodology

Given an example face image, a CNN network for face anti-spoofing application has a pre-defined task to output the probability of whether the given face image is spoofed or real. Suppose that the input face image x_i to a j_{th} CNN network is sampled randomly from a set S of face images, which contain both real and spoofed images. Then, the output of a j_{th} CNN network can be represented in abstract notation by using (1-3).

$$y_{ij} = f(\mathbf{CNN}(x_i \in S)_j), \quad (1)$$

$$f(input_i) = sofmax(input_i), \quad (2)$$

$$softmax(input_i) = \frac{e^{input_i}}{\sum_N e^{input_i}}. \quad (3)$$

In other words, (3) gives us the probability of an input face image x_i being fake or real after passing it through a j_{th} CNN architecture. In the proposed approach, we consider various CNN architecture based on conventional VGG-11 model. As the available training data is limited, and the network is deep, therefore we employ three different architectures of CNN with hyper-parameters like changing the resolution of first two layers of the CNN architecture and the addition of dropout etc. Addition of dropout in a network with limited data reduces the effect of overfitting as proposed in [25]. We also perform our experiments with changing the first two layers of the conventional VGG-11 network with and without dropout, and thus the resolution of weight layer become another hyper-parameter. In the following paragraphs, we provide the details on data preparation and training the proposed CNN model for face anti-spoofing application.

A. Data Preparation and Preprocessing

Before training a CNN architecture for face anti-spoofing on a given dataset, the images in the dataset are pre-processed, which usually includes mean centering and normalization. We utilize CASIA-FASD dataset for our extensive experiments and evaluation of CNN for face anti-spoofing. Since, CASIA-FASD dataset mainly consists of video-data rather than single images, the proposed data preparation process is different from conventional CNN data-preparation approaches for face anti-spoofing. In the conventional CNN approaches for face anti-spoofing, only a limited portion of video-data restricted to 10 to 20 frames of each video were utilized as an input data to the CNN architecture. On contrary, we randomly select 100 frames from each video and stored them in a disk. At training time, we detect the face area using voila-jones cascade classifier in the mini-batch image frames and normalize it before giving as an input to a CNN network. This process is shown in Fig. 1. In total, we form a training dataset that contains 24000 training images and 36000 test images. The reason for having

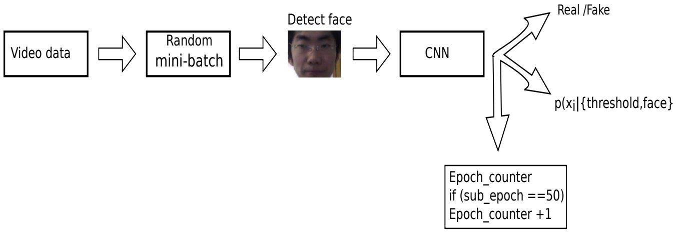

Fig. 1. Model structure: The proposed model consist of face detector with a deep CNN followed by two outputs: a binary classifier, and output probability of true class. An epoch counter is used to count sub-epochs

a low number of training images compared to test images is because of the limited subjects that are provided in the CASIA-FASD training dataset, i.e. 20 subjects in training dataset with 3 videos of real-face and 9 videos of fake-face per subject. The number of training images could be increased to a more higher number by performing data-augmentation on each image that include horizontal and vertical flipping, addition of random noise, horizontal and vertical translation. However, in the proposed case no data-augmentation has been performed.

B. Training

We use 50RS-30SeC-1E (50 Random Samples-30 Sub-epochs Count-1Epoch) strategy to train a VGG-11 network and its derived networks. For each forward-pass through the CNN network, we randomly sample 50 face images form the training dataset. Each forward-pass through a CNN network correspond to a single sub-epoch, and 30 sub-epochs count for 1 epoch. This process provides the advantage that at each foward-pass, the network is introduced with a new randomly sampled mini-batch. We trained our network for around 1000 epochs. If combined with sub-epochs, this make the total number of iterations to be 30,000. Training in this fashion provide certain advantages: First, the network can be trained on more than a single video-frame. That said, the network is then able to learn more about the discriminative features in the input data. Second, training on a known indexed single frame from the input video-data can make the network deterministic regarding the input data and may lead to overfitting. Thus, providing more frames from the video-data can circumvent the problem of overfitting. Also, the main reason for our network to be trained for more epochs is because at every forward-pass during training the CNN network is introduced with a random combination of data (that may include new video frames not known to CNN network) in the form of mini-batches, which provide new information to the network at every forward-pass and make the network to learn more about the inter-variation in the face images.

A VGG-11 Net has been used as our primary network A, and other networks B and C are derived from VGG-11 with some modifications as shown in Table I. For all CNN networks, the learning rate has been initially set to 0.01 for the first 100 epochs. The learning rate is then reduced by a factor of 0.1 for the next 100, 200 and 50 epochs respectively. Overall, we reduce the learning rate three times. The weight decay has been set to 0.005 with a dropout of 0.5 in the fully-connected layers for Network A, B and C, and an additional dropout of 0.5 after every max pooling layer in Network C. We use the Rectified Linear Unit(ReLU) as an activation functions for all the networks. The frames are resized to 96 x 96 patch keeping the center part of the patch and respecting the aspect-ratio. This is done by considering our system limitations and GPU capabilities.

TABLE I
VGG-11 AND ITS DERIVED NETWORKS

Network A	Network B	Network C
$96 \times 96 \times 3$		
3 x 3, 64 ReLU	7 x 7, 64 ReLU	7 x 7, 64 ReLU
2 x 2 max-pool	2 x 2 max-pool	2 x 2 max-pool, Dropout 0.5
3 x 3, 128 ReLU	5 x 5, 128 ReLU	5 x 5, 128 ReLU
2 x 2 max-pool	2 x 2 max-pool	2 x 2 max-pool ,Dropout 0.5
3 x 3, 256 ReLU	3 x 3, 256 ReLU	3 x 3, 256 ReLU
3 x 3, 256 ReLU	3 x 3, 256 ReLU	3 x 3, 256 ReLU
2 x 2 max-pool	2 x 2 max-pool	2 x 2 max-pool, Dropout 0.5
3 x 3, 512 ReLU	3 x 3, 512 ReLU	3 x 3, 512 ReLU
3 x 3, 512 ReLU	3 x 3, 512 ReLU	3 x 3, 512 ReLU
2 x 2 max-pool	2 x 2 max-pool	2 x 2 max-pool,Dropout 0.5
3 x 3, 512 ReLU	3 x 3, 512 ReLU	3 x 3, 512 ReLU
3 x 3, 512 ReLU	3 x 3, 512 ReLU	3 x 3, 512 ReLU
2 x 2 max-pool	2 x 2 max-pool	2 x 2 max-pool,Dropout 0.5
FC-4096 ReLU		
FC-4096 ReLU		
FC-2 ReLU		
Soft-max		

IV. EXPERIMENTAL RESULTS

After training CNN architectures, we perform extensive evaluation on CASIA-FASD face anti-spoofing dataset for performance evaluation using various metrics proposed in the literature. This include evaluation of top-1 percent accuracy, accuracy using threshold-operation, Equal Error Rate (EER), Quality Test (QT), Fake Image Test (FFT) and Overall Test (OT).

A. Results on CASIA-FASD

We train our algorithm first on CASIA-FASD database. The database has 50 real subjects with their spoofed faces generated using wrapped photo, cut-photo and video attacks respectively. Three imaging qualities are considered, i.e. low quality, normal quality and high quality. For each subject, the database provides 12 videos that contain three real faces and 9 fake-faces with a combination of above-mentioned attacks and imaging qualities.

Top-1 percent accuracy: In this method, only the output class with a highest probability is considered. The classes at the output of a CNN are arranged in a sequential manner, such that the index of the class corresponds to the label of the class. If the index of a class, having maximum probability among other class, matches with the corresponding true label the output counts 1 for the true positive or true negative otherwise the output counts 0.

Accuracy using threshold-operation: In this method, only a true (live face) class probability is considered. The probability

TABLE II
ACCURACY AND EER RESULTS ON CASIA-FASD DATABASE FOR VARIOUS PROPOSED CNN CONFIGURATION AND OTHER STATE OF THE ART METHODS

Network	Average Test Accuracy	Average Test Accuracy	EER
	Top-1 percent	Threshold-operation	
Network A	84%	85%	7%
Network B	81%	84%	7%
Network C	88%	89%	**5%**
CNN + SVM [23]	-	-	7.49%
Multi-Cues integration + NN [4]	-	-	5.83%

TABLE III
CLASSIFICATION ACCURACY FOR QT FOR THE PROPOSED CNN CONFIGURATIONS

Network	Normal	Low	High
Network A	91%	84%	79%
Network B	93%	86%	80%
Network C	94%	94%	82%

of the true class obtained at the output of a CNN network is compared against a set of threshold values to determine whether the output probability of the true class is higher than the corresponding threshold. If the output probability of a true class is higher than the threshold, the class is considered as true otherwise fake. The number of true positive, true negative, false positive and false negative at the corresponding threshold values are then determined by comparing the ground truth labels against the corresponding output labels obtained from CNN network.

The result of using top-1 percent accuracy has been show in Table II. As we can see from Table II, the Network C obtains an average top-1 percent test accuracy of 88%. In top-1 percent accuracy, only the output class with a highest probability among the other classes is considered. Since in network C, additional regularization has been done by utilizing dropout after every max-pooling layer, the amount of overfitting has been reduced in that network. The results obtained by using top-1 percent accuracy has been generated in the following way: For each input image the CNN will output class probabilities by using (4).

$$[p_1, p_2] = softmax(F_i). \tag{4}$$

where p_i is the probability that the input face image F_i belong to the true class and vice-versa. We use argmax function to get the indices of the largest probability. If the label of the true class matches with the indices of the class having largest probability, the output has been considered as either true positive or true negative, otherwise the output will be either false positive or false negative.

The second approach we utilize considers the probability of only true class (live face). In this approach the probability of the true class, after the soft-max layer as shown in Fig. 1, has been compared against a set of threshold values as proposed in traditional face anti-spoofing techniques [26], i.e. we compute, whether the probability of the true class is greater than the given threshold, and for the false class (fake face), whether the probability of the true class is less than the given threshold. We use a set of 100 threshold values from 0.01 to 1 with a difference of 0.01. After obtaining False Acceptance Rate (FAR) and False Rejection Rate (FRR) at various threshold values, the Equal Error Rate (EER) is calculated by using (5, 6).

$$T = \underset{T}{argmin}(|FAR - FRR|), \tag{5}$$

$$EER = \frac{FAR_T + FRR_T}{2}. \tag{6}$$

The result obtained by considering only the true class probability has been reported in Table II along with corresponding EER values under the label of Threshold-operation. It can be observed from Table II, that the best performance has been achieved by Network C, which is a regularized version of VGG-11 network. With Network C, an average test accuracy of 89% has been achieved with the lowest EER of 5% as compared to the state of the art approaches presented in [23], [4]. Additionally, the complex pre-processing steps performed in [4] like obtaining Face OFM map, Scene OFM maps are being performed internally by CNN architecture. The only pre-processing we perform include, the mean centering and normalization while the rest of the processes, i.e. learning features and correspondingly mapping face images to corresponding class probabilities have been performed by CNN by utilizing back propagation.

B. Protocol Test on CASIA-FASD

We also evaluate the proposed CNN model on the test protocol provided in [2], [4]. The protocol list seven scenarios that can be divided into three tests: Quality Test (QT), Fake Face Test (FFT) and Overall Test(OT).

1) Quality Test (QT): This test is used to evaluate the performance of a face anti-spoofing system given that the input image quality is fixed. The samples used to perform evaluation are mentioned as:

1. Low(L) quality test $L1, L2, L3, L4$
2. Normal (N) quality test: $N1, N2, N3, N4$
3. High (H) quality test: $H1, H2, H3, H4$

2) Fake Face Test (FFT): This test is used to evaluate the performance of a face anti-spoofing system given that fake face types are fixed. The samples used to perform the evaluation are mentioned as:

1. Wrapped photo attack test: $L1, N1, H1, L2, N2, H2$
2. Cut photo attack test: $L1, N1, H1, L3, N3, H3$
3. Video attack test: $L1, N1, H1, L4, N4, H4$

TABLE IV
ACCURACY USING FFT FOR THE PROPOSED CNN CONFIGURATIONS

Network	Wrapped photo attack	Cut photo attack	Video attack
Network A	79%	80%	82%
Network B	80%	82%	82%
Network C	87%	87%	86%

TABLE V
ACCURACY OF CNN VS [26] ON QT

Technique	Low	Normal	High
LBP+SVM	78%	83%	90%
LBP + SAE	75%	83%	90%
SBFD + SVM	91%	89%	82%
SBFD + SAE	94%	92%	87%
CNN (Ours)	94%	94%	82%

3) Overall Test (OT): This test is used to evaluate the performance of face anti-spoofing system by combining all the data to perform a general evaluation.

Table III list the corresponding test accuracies of the proposed three networks for QT test. As can be observed in Table III, Network C (which is a VGG-11 network with addition of dropout after every max pooling layer) gives an overall high classification rate as compared to remaining CNN networks. Similarly, for FFT test as shown in Table IV, Network C provided an overall high accuracy of above than 82% for all three types of attacks. Since in the previous work [23], a CNN is mainly used for feature extraction and SVM for classification; keeping in view the recent success in image classification [7], we utilize a CNN architecture to directly classify an input face into either live-face or fake-face. Table V provides a comparative results for QT test by using the proposed CNN model for face anti-spoofing against other state of the art face anti-spoofing approaches [26] that include Local Binary Patterns + Support Vector Machine (LBP + SVM), Local Binary Patterns + Stacked Auto Encoders (LBP + SAE), Shearlet Based Feature Descriptors + Support Vector Machine (SBFD + SVM) and Shearlet Based Feature Descriptors + Stacked Auto Encoders (SBFD + SAE). In Table V and subsequently in the rest of the tables, we only report the accuracies obtained by using our best performing model, i.e. Network C. Table VI reports the accuracies obtained on FFT test for the proposed CNN model and other state of the art face anti-spoofing techniques. As can be seen in Table VI, accuracies obtained on FFT test for face anti-spoofing using CNN model is almost on par with other corresponding state of the art techniques. Table VII compares the results of our best performing CNN network for face anti-spoofing with other state of the art face anti-spoofing methods for OT test. As can be seen in Table VII, a 3% improvement has been achieved in liveness detection using the proposed CNN model. Additionally, in Table V-VII, the proposed CNN model for face anti-spoofing application perform almost on par with other state of the art approaches proposed in [26]. However, in CNN models, we dont need the hard process of feature extraction and features selection for getting the better accuracies for various face anti-spoofing scenarios. Additionally, in Table VII, we also reported the OT test for multi-class classification (Liveness + attacks classification) besides binary classification (Liveness detection). The results of Table VII show that the proposed CNN architecture perform on par with the approach proposed in [26] for face anti-spoofing application. However for liveness detection problem, the result on OT using the proposed CNN model is 3% better than the approach of [26] that use handcrafted feature and stacked auto-encoder.

Compared to the traditional approaches that used fixed handcrafted features, in the proposed work, we train our CNN models using end-to-end learning for feature learning, i.e. we only input the face images to the network and let the network learn the features from the face images. The only pre-processing we performed is the mean centering and normalization on the input data. More importantly, for training a CNN model, we dont need any developmental data for obtaining a best threshold values for evaluating the CNN model performance for face anti-spoofing application. Additionally, to further explore the capabilities of CNN for face anti-spoofing and to set our future directions, we could use more advanced CNN architectures and more complex face anti-spoofing datasets and protocols.

TABLE VI
ACCURACY OF CNN VS [26] ON FFT

Technique	Wrapped photo attack	Cut photo attack	Video attack
LBP + SVM	82%	79%	80%
LBP + SAE	85%	83%	90%
SBFD + SVM	83%	92%	92%
SBFD + SAE	83%	93%	91%
CNN (Ours)	87%	87%	86%

TABLE VII
ACCURACY OF CNN VS [26] ON OT

Technique	Multi-Class Classification	Liveness Detection
SBFD	80.86%	89.18%
CNN (Ours)	80.91%	92.52%

V. CONCLUSION

In this paper, we proposed a CNN framework for face anti-spoofing application. An end-to-end learning approach, where the input to a CNN is a face image and the output is class probability, has been proposed. Additionally, an efficient training strategy has been proposed for training a CNN architecture when training data is limited. Experimental results clearly demonstrate the effectiveness of the proposed CNN architecture on both top-1 percent accuracy and traditional performance evaluation metrics. This also provides a platform to assess further the capabilities of various CNN paradigms using end-to-end learning approach on other face anti-spoofing databases.

Acknowledgment

The work described in this paper was substantially supported by Internal SRG Funding from City University of Hong Kong under the Project No. 7004430.

References

[1] Z. Boulkenafet, Z. Akhtar, X. Feng, and A. Hadid, "Face anti-spoofing in biometric systems," in *Biometric Security and Privacy*. Springer, 2017, pp. 299–321.

[2] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face antispoofing database with diverse attacks," in *Biometrics (ICB), 2012 5th IAPR international conference on*. IEEE, 2012, pp. 26–31.

[3] R. Ramachandra and C. Busch, "Presentation attack detection methods for face recognition systems: a comprehensive survey," *ACM Computing Surveys (CSUR)*, vol. 50, no. 1, p. 8, 2017.

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

[5] D. Menotti, G. Chiachia, A. Pinto, W. R. Schwartz, H. Pedrini, A. X. Falcao, and A. Rocha, "Deep representations for iris, face, and fingerprint spoofing detection," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 864–879, 2015.

[6] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[13] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European Conference on Computer Vision*. Springer, 2016, pp. 21–37.

[16] M.-A. Waris, H. Zhang, I. Ahmad, S. Kiranyaz, and M. Gabbouj, "Analysis of textural features for face biometric anti-spoofing," in *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*. IEEE, 2013, pp. 1–5.

[17] J. Määttä, A. Hadid, and M. Pietikäinen, "Face spoofing detection from single images using micro-texture analysis," in *Biometrics (IJCB), 2011 international joint conference on*. IEEE, 2011, pp. 1–7.

[18] ——, "Face spoofing detection from single images using texture and local shape analysis," *IET biometrics*, vol. 1, no. 1, pp. 3–10, 2012.

[19] J. Galbally, S. Marcel, and J. Fierrez, "Biometric antispoofing methods: A survey in face recognition," *IEEE Access*, vol. 2, pp. 1530–1552, 2014.

[20] D. Gragnaniello, C. Sansone, G. Poggi, and L. Verdoliva, "Biometric spoofing detection by a domain-aware convolutional neural network," in *Signal-Image Technology & Internet-Based Systems (SITIS), 2016 12th International Conference on*. IEEE, 2016, pp. 193–198.

[21] A. Alotaibi and A. Mahmood, "Deep face liveness detection based on nonlinear diffusion using convolution neural network," *Signal, Image and Video Processing*, pp. 1–8, 2016.

[22] ——, "Enhancing computer vision to detect face spoofing attack utilizing a single frame from a replay video attack using deep learning," in *Optoelectronics and Image Processing (ICOIP), 2016 International Conference on*. IEEE, 2016, pp. 1–5.

[23] J. Yang, Z. Lei, and S. Z. Li, "Learn convolutional neural network for face anti-spoofing," *arXiv preprint arXiv:1408.5601*, 2014.

[24] 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 *Image Processing Theory Tools and Applications (IPTA), 2016 6th International Conference on*. IEEE, 2016, pp. 1–6.

[25] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[26] L. Feng, L.-M. Po, Y. Li, and F. Yuan, "Face liveness detection using shearlet-based feature descriptors," *Journal of Electronic Imaging*, vol. 25, no. 4, pp. 043 014–043 014, 2016.