Attention-Based Two-Stream Convolutional Networks for Face Spoofing Detection

Chen, H., Hu, G., Lei, Z., Chen, Y., Robertson, N., & Li, S. (2019). Attention-Based Two-Stream Convolutional Networks for Face Spoofing Detection. *IEEE Transactions on Information Forensics and Security*. Advance online publication. https://doi.org/10.1109/TIFS.2019.2922241

Published in:
IEEE Transactions on Information Forensics and Security

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Attention-Based Two-Stream Convolutional Networks for Face Spoofing Detection

Haonan Chen[1*], Guosheng Hu[2*], Zhen Lei[3], Yaowu Chen[1△], Neil M. Robertson[2] and Stan Z.Li[3]

[1]Zhejiang Provincial Key Laboratory for Network Multimedia Technologies, Zhejiang University
[2]Queens University Belfast
[3]Institute of Automation, Chinese Academy of Sciences

***Abstract*—Since the human face preserves the richest information for recognizing individuals, face recognition has been widely investigated and achieved great success in various applications in the past decades. However, face spoofing attacks (e.g. face video replay attack) remain a threat to modern face recognition systems.Though many effective methods have been proposed for anti-spoofing, we find that the performance of many existing methods is degraded by illuminations. It motivates us to develop illumination-invariant methods for anti-spoofing. In this paper, we propose a two stream convolutional neural network (TSCNN) which works on two complementary space: RGB space (original imaging space) and multi-scale retinex (MSR) space (illumination-invariant space). Specifically, RGB space contains the detailed facial textures yet is sensitive to illumination; MSR is invariant to illumination yet contains less detailed facial information. In addition, MSR images can effectively capture the high-frequency information, which is discriminative for face spoofing detection. Images from two spaces are fed to the TSCNN to learn the discriminative features for anti-spoofing. To effectively fuse the features from two sources (RGB and MSR), we propose an attention-based fusion method, which can effectively capture the complementarity of two features. We evaluate the proposed framework on various databases, i.e. CASIA-FASD, REPLAY-ATTACK and OULU, and achieve very competitive performance. To further verify the generalization capacity of the proposed strategies, we conduct cross-database experiments, and the results show the great effectiveness of our method.**

***Index Terms*—Face spoofing, multi-scale retinex, deep learning, attention model, feature fusion.**

I. Introduction

COMPARED with traditional authentication approaches including password, verification code and secret question, biometrics authentication is more user-friendly. Since the human face preserves rich information for recognizing individuals, face becomes the most popular biometric cue with the excellent performance of identity recognition. Currently, person identification can easily use the face images captured from a distance without physical contact with the camera on the mobile devices, e.g. mobile phone.

As the application of face recognition system becomes more and more popular with the widespread of the Mobile phone, their weaknesses of security become increasingly conspicuous. For example, owing to the popularity of social network, it is quite easy to access a person's face image on the Internet to attack a face recognition system. Hence, a deep attention for face spoofing detection has been drawn and it has motivated great quantity of studies in the past few years.

In general, there are mainly four types of face spoofing attacks: photo attack, masking attack, video replay attack and 3D attack. Due to the high cost of the masking attack and 3D attack, therefore, the photo attack and video reply attack are the two most common attacks. Photo and video replay attacks can be launched with still face images and videos of the user in front of the camera, which are actually recaptured from the real ones. Obviously, the recaptured image is of lower quality compared with the real one in the same capture conditions. The lower quality of attacks can result from: lack of high frequency information [1]–[5], image banding or moire effects [6], [7], video noise signatures, etc. Clearly, these image quality degradation factors can work as the useful cues to distinguish the real faces and the fake ones.

Face spoofing detection, which is also called face liveness detection, has been designed to counter different types of spoofing attacks. Face spoofing detection usually works as a preprocessing step of the face recognition systems to judge whether the face image is acquired from a real person or a printed photo (replay video). Therefore, face spoofing detection is actually a binary classification problem.

To counter the face spoofing attacks, there are mainly four solutions available in the research literature: (1) micro-texture based methods, (2) image quality based methods, (3) motion based methods, and (4) reflectance based methods. For (1), local micro-texture features are demonstrated as a useful cue when attacked by photo and video. Researchers start the texture-based methods by feeding hand-crafted features extracted from facial texture to classifiers [8]–[12]. With the development of deep learning, CNN [13]–[15] is utilized to learn discriminative features for face spoofing detection. For (2), the low imaging quality of the fake images offers the useful clues [1]–[7], e.g. the loss to high frequency information, these clues have successfully been used for spoofing detection. For (3), motion-based methods mainly contain: physiological reaction based [16]–[18] and physical movement based [19], [20]. Motion-based methods may become less effective when conducted by video replay which can present the facial motions. For (4), reflectance of the face image is another widely used cue for liveness detection because the lighting reflectance from real face (3D) and attacking (mostly

*Equal Contribution
△ Corresponding author (cyw@mail.bme.zju.edu.cn)

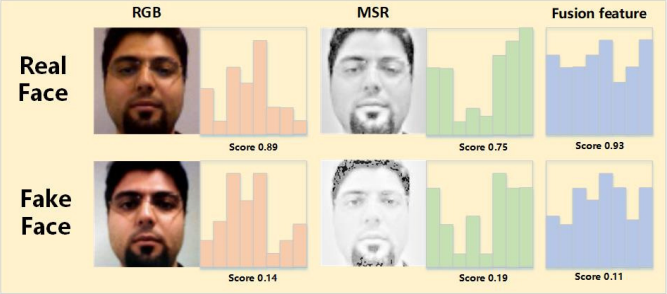


Fig. 1. Motivation of the fusion of RGB (*Col 1*) and MSR (*Col 3*) images. The individual feature scores of RGB (*Col 2*) and MSR (*Col 4*) and fused scores (*Col 5*) are shown. The fused scores are improved compared with individual scores.

2D, such as photo and replay attacks) face is very different [1], [21], [22].

In this work, we propose a novel deep learning based micro-texture based (MTB) method. The existing MTB methods usually process and analyze the input images in original RGB color space. However, the RGB images are sensitive to illuminations. The RGB based MTB methods can potentially reduce their performance in the presence of illuminations. This motivates us to develop a illumination-robust MTB method. Therefore, we proposed a two-stream convolutional neural network (TSCNN) which is trained on two complementary space: RGB space (original space) and multi-scale retinex (MSR) [23] space (illumination-invariant space).

First, both RGB and MSR images contain discriminative information: RGB images can be used to train end-to-end discriminative CNNs for spoofing detection; MSR can capture high frequency information, and this information is verified particularly effective for spoofing detection. Second, RGB and MSR images are complementary: RGB space contains the detailed facial information yet is sensitive to illumination; MSR is invariant to illumination yet contains less detailed facial information. In the framework of TSCNN, the RGB and MSR images are fed to two CNNs (two branches of TSCNN) separately and generate two features which are discriminative for anti-spoofing. To effectively fuse these two features, we propose a learning-based fusion method inspired by attention mechanism [24] detailed in Section III-C. Apart from the commonly used fusion methods, e.g. feature averaging fusion, our attention-based fusion can adaptively weight features to achieve promising performance of fused features. Fig.1 shows the complementarity of RGB and MSR and the importance of the feature fusion. Our contributions can be summarized as:

- We propose a two-stream CNN (TSCNN) which accepts two complementary information (RGB and MSR images) as input. To our knowledge, we are the first to investigate the fusion of these two discriminative clues (RGB and MSR) for face anti-spoofing.
- To adaptively and effectively fuse two features generated by TSCNN, we proposed an attention-based fusion method. The proposed fusion method can make the TSCNN generalize well to images under various lighting conditions.
- We conduct extensive evaluations on three popular anti-spoofing databases: CASIA-FASD, REPLAY-ATTACK and OULU. The results show the effectiveness of the proposed strategies. In addition, we run cross-database experiments with very competitive results, showing the great generalization capacity of the proposed method.

II. Related Works

A. Face Spoofing Detection

In these years, various methods have been proposed for face spoofing detection. In this section, we briefly review the existing anti-spoofing methods.

Texture Based Methods Texture based methods focus on exploring different texture-based features for face spoofing detection. The features can be simply classified as: hand-crafted features and deep learning based features.

We first introduce hand-crafted feature based method. Based on the idea that specific frequency bands preserve most texture information of real faces, the work in [3] employed various difference-of-Gaussian filters to select a favorable frequency band for detection. Texture features used in face detection and face recognition tasks can be migrate to face spoofing detection and perform quite well.

Apart from hand-crafted features, deep learning, in particular, CNN based features achieved great success in recent years. In this category, the CNN learns the discriminative features for liveness detection. The large amount of training data guides the CNN to learn an effective feature. [25] extracts the local texture features and depth features from the face images and fuses them for face spoofing detection. Furthermore, a LSTM-CNN architecture [26] was proposed to fuse the predictions of the multiple frames of a video, which was proved to be effective for video face spoofing detection.

Image Quality Based Methods Methods in this category are motivated by the fact that the photo and replay video are likely to have an image quality degradation in the recapture process. In [1], the method exploits to analyze the attack photos in 2D Fourier spectra, showing interesting results. However, the performance might drop for higher-quality image data. Moreover, in [5], an image quality based method was proposed by applying chromatic moment feature, specular reflection feature, blurriness feature and color diversity feature.

Motion Based Methods This type of methods aim to select the physiological reaction motions such as eye blinking, lips movements and the head motions to distinguish the real face from the fake one. In [20], different movements in the facial parts were extracted as features for this task. Though physiological sign based methods have shown satisfactory performance to counter printed photo attacks with the user cooperation, they may become less effective for video replay attack. However, [27] advances a method for facial anti-spoofing by applying dynamic mode decomposition (DMD), which can conveniently represent the temporal information of the replay video as a single image with the same dimensions as frames in the video. This method based on the motion information is proved less time consuming and is more accurate.

Reflectance Based Methods The reflectance differences between the real and fake faces, in particular for the print

attack and replay attack, can offer important information for face spoofing detection. The reflectance cue from a single image is used to detect the face spoofing [1], [22]. [28] utilizes the different multi-spectral reflectance distributions to distinguish real and fake faces based on Lambertian model.

Multi-Feature Fusion Based Methods The fusion of multiple features show improved accuracy compared to individual feature. [29] proposed a feature fusion with video motion feature and texture feature to distinguish the authenticity of the face. The author obtains the moving image from the face video and the LBP feature from the last frame, fuses them and uses the linear discriminant analysis (LDA) for classification. [9] extracts the texture features from three multi-scale filtering methods, then the resulting features are concatenated to form the fused feature for classification.

Other Methods Apart from the aforementioned methods, additional hardwares can also be employed for face spoofing detection. Unlike face images directly captured by camera, 3D depth information [30]–[32] and multi-spectrum and infrared (IR) image. [30] proposed a method for face liveness detection based on 3D projective invariants. In [31], the authors proposed to recover sparse 3D shapes for face images to counter the different kinds of photo attacks.

Summary The methods we introduced can usually achieve promising performance of anti-spoofing on intra-database scenario, however, it is still challenging to achieve strong performance for inter-database scenario. The degraded generalization capacity results from many cross-database factors: different capture devices, different imaging environments, different illuminations, different facial poses, etc. In this work, we propose an anti-spoofing method which is illumination-robust, generalizing well to environments with strong illumination environments and without, achieves promising cross-database performance.

B. Multi-Scale Retinex

Many related researches have been conducted to simulate the human vision system using different luminance algorithms. Land's Retinex theory [33] proposed the a lightness model named as Retinex theory to measure the lightness reflexion in an image. After that, the Retinex algorithm has been successfully applied to image enhancement [34], [35]. [36] introduced a model called Single Scale Retinex (SSR), which applied the Gaussian filter to normalize illumination of source image. The work [37] focused on the filter of the SSR and employed an improved SSR with the guided filter and achieved promising image enhancement performance. The performance of SSR algorithm is highly dependent on the parameter of Gaussian filter. To overcome this limitation, a multi-scale Retinex (MSR) model [23], which weights the outputs of several SSRs, is proposed. [38] proposed a novel MSR based on an adaptive weights to aggregate the SSRs and applied in image contrast enhancement. In our work, we applied MSR because: (1) MSR can separate an image to illumination component and reflectance component, and the illumination-removed reflectance component is used for liveness detection; (2) the MSR algorithm can be regarded as a optimized high pass filter, thus it can effectively preserve the high frequency components which is discriminative between the real and fake faces.

C. Feature Fusion

Existing fusion methods consist of two part: early fusion (feature-level fusion) and late fusion (Score-level fusion). Feature aggregation or subspace learning is actually the early fusion. Aggregation approaches are usually performed by simply element averaging or concatenation [39]. Subspace learning methods aim to project the concatenated feature to a subspace with the best use of the complementarity of the features. Late fusion is to fuse the predicted scores after computation based on different classifier by averaging [40] or stacking another classifier result [41]. For the deep learning task, researchers usually use simple fusion methods for fusing deep features features, such as score fusion, feature averaging, etc. In our work, we proposed an attention based fusion method, aiming to exploit the best use of the features to fuse.

D. Visual Attention Model

Visual attention is a powerful mechanism that enables perception to focus on important part which offers more information. To combine spatial and temporal information [42] employed an end-to-end deep neural network. In [43], the authors proposed a novel visual attention model to integrate different spatial features including color, orientation and luminance orientation features, which can reflect the region of interests of the human visual system. Different mechanisms of attention have been employed to deal with the computer vision tasks, including action recognition [44], emotion recognition [45], image classification [46]. On the whole, the attention model is usually used for aggregating features extracted by different images. Inspired by the great success of attention models, we apply attention model to fuse our features derived from RGB images and MSR images.

III. Methodology

Spoofing detection is actually a binary (real vs. fake face) classification problem. In deep learning era, a natural solution of this task is to feed the input RGB images to a carefully designed CNN with classification loss (softmax and cross entropy loss) for end-to-end training. This CNN-based framework has been widely investigated by [25], [26], [47]–[50].

Despite the strong nonlinear feature learning capacity of deep learning, the performance of anti-spoofing degrades when the input images are captured by different devices, under different lighting, etc. In this work, we aim to train a CNN which generalizes better to various environments, mainly various lightings.

The RGB images are sensitive to illumination variations yet cover very detailed facial texture information. Motivated by extensive research of (single-scale and multi-scale) Retinex image, we find the Retinex (we use Multi-Scale Retinex - MSR in this work) image is invariant to illumination yet loses minor facial texture. Thus, in this work, we propose a

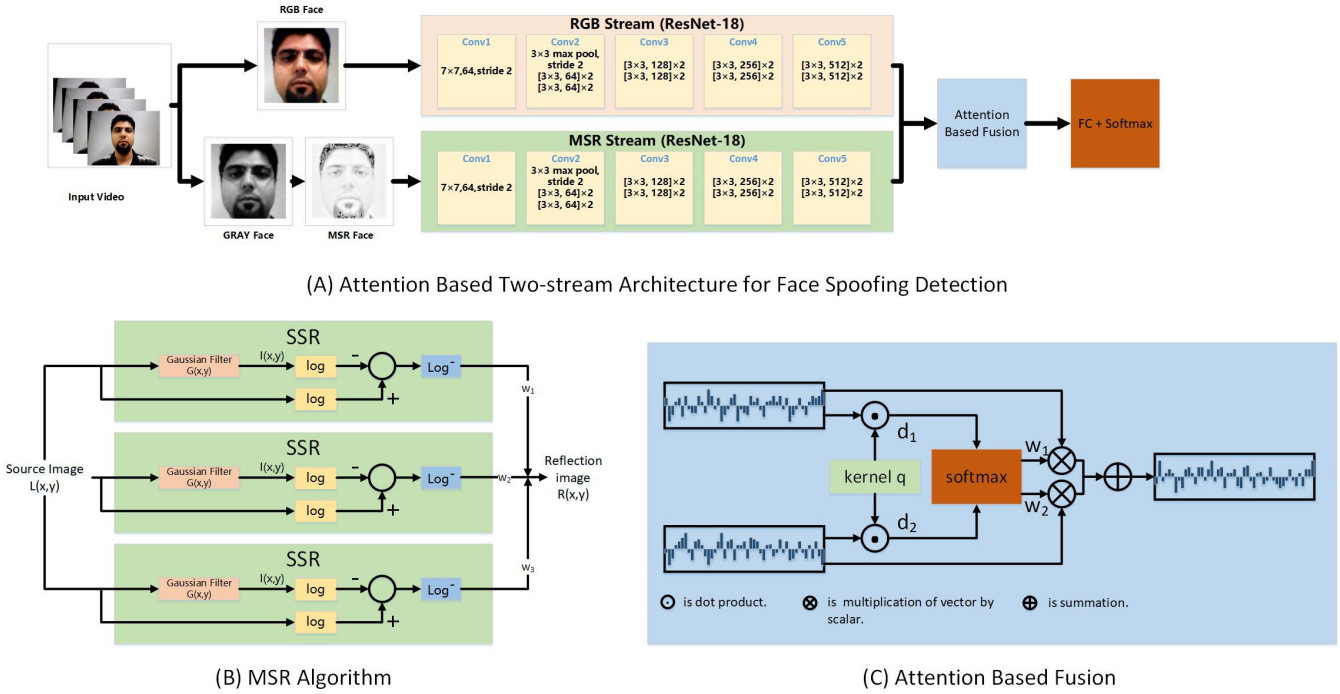


Fig. 2. (A) is the overall pipeline; In (B), every single block represents one SSR module. The outputs of all SSR modules are weighted with scale parameters to form MSR; (C) illustrates the work flow of attention-based fusion.

two-stream CNN (TSCNN) which trains two separate CNNs accepting RGB images and MSR images as input respectively. To effectively fuse RGB feature and MSR feature, we propose an attention based fusion method.

In this section, firstly, we introduce the theory of the Retinex to explain the reason why MSR image is discriminative for anti-spoofing. After that, the complementarity of the RGB and MSR features is analyzed and the proposed TSCNN is detailed. Last, we introduce our attention-based feature fusion method.

A. The Retinex Theory

Assumption Retinex theory was first raised by Land and McCann in 1971 [33]. According to the literal meaning of the word 'Retinex', it is a portmanteau constituted by 'retina' and 'cortex', imitating how the human visual system works. The Retinex theory is based on the assumption that the color of the object is determined by the reflection ability of light of different wavelengths. The color of the object is not affected by the non-uniformity illumination. The theory separates the source image $S(x,y)$ into two parts: the reflectance $R(x,y)$ and the illumination $L(x,y)$. In particular, $R(x,y)$ and $L(x,y)$ contain different components of frequency. $R(x,y)$ focuses on high frequency components, while $L(x,y)$ tends to low frequency components. We formulate Retinex by Eq. (1):

$$S(x,y) = R(x,y) \cdot L(x,y) \tag{1}$$

where x and y are image pixel coordinates.

Motivation $L(x,y)$ and $R(x,y)$ represent the illumination and reflectance (facial skin texture in our task) components respectively. $L(x,y)$ is determined by the light source, while $R(x,y)$ is determined by the property of the surface of captured objects, i.e face in our application. Illumination is clearly not relevant to most classification tasks including face spoofing detection, thus the separation of illumination and reflectance (texture) is important because the separated reflectance only can be used for illumination-invariant classification. Since Retinex theory aims to conduct this separation, Retinex is used in this work for illumination-invariant face spoofing detection.

Computation For the convenience of calculation, Eq. (1) is usually transformed into the logarithmic domain:

$$log[S(x,y)] = log[R(x,y)] + log[L(x,y)] \tag{2}$$

where $log[S(x,y)]$, $log[R(x,y)]$, and $log[L(x,y)]$ are represented by $s(x,y)$, $r(x,y)$, and $l(x,y)$ for convenience.

Since $s(x,y)$ is logarithmic form of the original image, we can calculate the Retinex output $r(x,y)$ by appraising $l(x,y)$. Thus, the performance of the Retinex is determined by the estimation of $l(x,y)$. Selecting the apposite method to estimate $l(x,y)$ is a considerable step for illumination normalization.

Summarizing the previous work of the Retinex, the illumination image can be generated from the source image using the center/surround Retinex. Single-scale Retinex (SSR) [36] is a center/surround based Retinex and is formulated as Eq. (3):

$$r(x,y) = s(x,y) - log[S(x,y) * F(x,y)] \tag{3}$$

where $F(x,y)$ denotes the surround function, and Symbol '*' is the convolution operation. There are several forms of the surround function which depends on the effect of the SSR.

The work [36] shows that a Gaussian filter works well for the illumination normalization.

$$G(x,y) = Ke^{-(x^2+y^2)/c} \tag{4}$$

where c is the scale parameter of Gaussian surround function. The value of c is empirically determined. K is selected to satisfy:

$$\iint F(x,y)dxdy = 1 \tag{5}$$

Let $G(x,y)$ represent $F(x,y)$, then Eq. (3) can be rewritten as:

$$r(x,y) = s(x,y) - log[S(x,y) * G(x,y)] \tag{6}$$

The large illumination discontinuities produce halo effects which are often visible. This limitation expands SSR to a more balanced method, multi-scale retinex (MSR) [23], by superposing several outputs of SSRs with small, middle, and large scale parameters at certain weights, shown in Fig.2 (B). Specifically, this is expressed by,

$$r_{MSR}(x,y) = \sum_{i=1}^{k} w_i \left\{ log[S(x,y)] - log[S(x,y) * G_i(x,y)] \right\} \tag{7}$$

Summary Retinex (MSR in our work) is used for face spoofing detection with two reasons. (1) The MSR can separate illumination and reflectance. In this work, we use the reflectance images (MSR image) to train a CNN for illumination-invariant face spoofing detection. (2) Since the fake face image is regraded as the recaptured image in many cases, which may lose some high frequency information compared to genuine ones. Thus, high frequency information can work as a discriminative clue for anti-spoofing. MSR algorithm can be viewed as an optimized high pass filter to capture the high frequency information for spoofing detection.

B. Two Stream Convolutional Neural Network (TSCNN)

In this section, we introduce our framework for anti-spoof: TSCNN. Specifically, the original RGB images are converted to MSR images in an off-line way. The two image sources (RGB and MSR) are separately fed to two CNN for end-to-end training with cross-entropy binary classification loss. The learned two features (derived from RGB and MSR images) are then learned to fuse using attention mechanism. In the remaining parts of this section, we will detail each component of our framework.

Complementarity of RGB and MSR Images RGB color space is commonly used for capturing and displaying color images. The advantage of the use of RGB images is clear: RGB images can naturally capture detailed facial texture which is discriminative for spoofing detection. However, the disadvantage of RGB image is that it is very sensitive to illumination variation. The intrinsic reason is that RGB space has high correlation between the three color channels, making it rather difficult to separate the luminance and chrominance information. Because the luminance conditions of face images in real world are different and the separation of luminance (illumination) and chrominance (skin color) is rather difficult, the features learned from RGB space tend to be affected by illumination.

The MSR algorithm can achieve illumination invariant face image by removing the illumination effects as introduced in SectionIII-A. Thus, the MSR face image preserves the micro-texture information of facial skin without the illumination effects. Apart from the illumination-invariant merit of MSR images, MSR images can generate discriminative information for spoofing detection. Specifically, MSR algorithm removes the low frequency components (illumination) from the original image and leaves the high frequency ones (texture details). However, the high frequency information is discriminative for spoof detection because: the real faces have rich facial texture details, while the fake faces, in particular recaptured faces, lose some of such details.

As analyzed above, RGB and MSR images are complementary because: RGB images contain detailed facial texture yet are sensitive to illuminations; while MSR images contain less detailed texture yet are illumination invariant. In addition, MSR images can keep high frequency information, which is also discriminative for spoofing detection.

Two-stream Architecture Our method is motivated by the fact that both RGB and MSR features are discriminative for face spoofing detection. It is natural to train CNNs using these two sources of information. In this work, therefore, we proposed a two-stream convolutional neural network (TSCNN) as shown in Fig.2 (A). The TSCNN consists of two identical sub-networks with different inputs (RGB and MSR images) and extract the learned features derived from RGB and MSR images following the last convolution layer of the two sub-networks. Given one input image/frame, we use MTCNN [51] for face and landmark detection. Then the detected faces are aligned using affine transformation. The RGB stream operates on single RGB frames extracted from a video sequence. For the MSR stream, the single RGB frames (processed to gray scale first) are converted to MSR images as shown in Fig.2-(B). Then MSR images are fed to the MSR subnetwork for training. Each stream is based on the same network, in this work, we use two successful networks (MobileNet [52] and ResNet-18 [53]). To effectively fuse the features from two streams, we propose an attention based fusion block, shown in Fig.2-(C), which will be detailed in Section III-C.

To formulate the TSCNN framework (M), we introduce a quadruplet $M = (E_{RGB}, E_{MSR}, F, C)$. Here E_{RGB} and E_{MSR} are features extractors for RGB and MSR streams respectively. F is a fusion function and C is the classifier. The feature extractor is a mapping $E : I \rightarrow f$ that takes an input image (either RGB or MSR) I and outputs a feature f of D-dimension.

Both the extracted feature f_{RGB} and f_{MSR} must have the same dimension of D to be compatible for early (feature) fusion. In particular, f_{RGB} and f_{MSR} can be obtained via different extractors (CNNs), while the feature dimension should be assured the same.

The fusion function F aggregates f_{RGB} and f_{MSR} into a fused feature v via F:

$$v = F(f_{RGB}, f_{MSR}) \tag{8}$$

The fused feature is then fed into a classifier C. Thus, the TSCNN can be formulated as an optimization problem:

$$\min_{w} \frac{1}{N} \sum_{i=1}^{N} l[C(F(f_{RGB}, f_{MSR})), y] \tag{9}$$

where $l(:,:)$ is a loss function, N is the number of samples, y is the one-hot encoding label vector.

Backbone Deep Networks CNNs have been successfully applied to face anti-spoofing [25], [26], [47]–[49]. Most existing works trained their CNN models from scratch using the existing face anti-spoofing databases, which are quite small and captured in unitary environments. Since CNNs are data hungry model, small training data might lead to overfitting. To overcome overfitting and improve the performance of many computer vision tasks, model finetuning/pretraining from big image classification database, usually ImageNet [54], is an effective way. In this work, we used two backbone networks pretrained on ImageNet, i.e MobileNet [52] (lighter, less accurate) and ResNet-18 [53] (heavier, more accurate) for spoofing detection.

To adapt the MobileNet and ResNet-18 models to our face anti-spoofing problem, we finetuned the pretrained models using the face spoofing database. The 2-class cross-entropy loss, i.e. Eq (10), is used for binary classification (real vs fake faces). The output of bottleneck layers of MobileNet (1024D) and ResNet-18 (512D) models work as the features for anti-spoofing.

$$C = -\frac{1}{N} \sum_{i}^{N} [y_i ln\hat{y}_i + (1 - y_i) ln(1 - \hat{y}_i)] \tag{10}$$

where i is the index of training sample, N is the number of training samples, $\hat{y}_i$ is the predict value of the i_{th} sample, y_i is the label of the i_{th} sample.

C. Attention based Feature fusion

Feature fusion is important for performance improvement in many computer vision tasks. Improper fusion methods can make the fused feature works worse than individual features. In deep learning era, fusion methods including score averaging, feature concatenation, feature averaging, feature max pooling and feature min pooling are normally used. In our anti-spoofing task, we find these fusion methods cannot explore deeply the interplay of features from different sources, therefore, we propose an attention-based fusion method as shown in Fig.2-(C).

The proposed attention-based fusion methods is actually a general framework which can be used for many deep learning based fusion scenarios, certainly including the fusion of RGB and MSR features. Given a set of features $\{f_i, i = 1, ..., N\}$, we try to learn a set of weights corresponding to the features $\{w_i, i = 1, ..., N\}$ to generate the aggregated feature v:

$$v = \sum_{i=1}^{N} w_i f_i \tag{11}$$

Clearly, the key part of our attention method is to learn the weights $\{w_i\}$ of Eq. (11). Note that our method becomes feature average fusion if $w_i = 1/N$, showing the generalization capacity of our method. In our task of spoofing detection, $N = 2$, and the features to be fused are f_{RGB} and f_{MSR}.

Apart from learning w_i directly, we learn a kernel q which has the same dimensionality of f_i. q is used to filter the feature vectors via dot product:

$$d_i = q^T f_i \tag{12}$$

The filter generates a vector which represent the significance of the corresponding feature, named d_i. To convert the significances to weights w_i subject to $\sum_i w_i = 1$, we passed d_i to a softmax operator and achieve all positive weights w_i:

$$w_i = \frac{e^{d_i}}{\sum_j e^{d_j}} \tag{13}$$

Obviously, the aggregation result r is unrelated with the quantity of input feature f_i. The only parameters to learn is the filter kernel q, which is easy to be trained via standard backpropagation and stochastic gradient descent.

IV. EXPERIMENTS

In this Section, we conduct extensive experiments and evaluate our method. We first have a brief introduction of three benchmark databases in Section IV-A. After that, we present the experimental settings of our method in section B so that the other researchers can reproduce our results. The following sections (SectionIV-C to G) present the results on the three databases. In particular, the results on CASIA-FASD are shown with the seven test scenarios.

A. Benchmark Database

In this subsection, to assess the effectiveness of our proposed anti-spoofing technique, an experimental evaluation on the CASIA Face Anti-Spoofing Database [55], the REPLAY-ATTACK database [56] and the OULU database [57] is provided. These three datasets consist of real client accesses and different types of attacks, which are captured in different imaging qualities with different cameras. In the following paragraphs, we will have a brief introduction of the databases.

1) The CASIA Face Anti-Spoofing Database (CASIA FASD): The CASIA Face Anti-Spoofing Database is divided into the training set consisted of 20 subjects and the test set containing 30 individuals(see, Fig.3). The fake faces were made by capturing the genuine faces. Three different cameras are used in this database to collect the videos with various imaging qualities: low, normal, and high. In addition, the individuals were asked to blink and not to keep still in the videos to collect abundant frames for detection. Three types of face attacks were designed as follows: 1) Warped Photo Attack: A high resolution (1920 $\times$ 1080) image, which is recorded by a Sony NEX-5 camera, was used to print a photo. The attacker simulates the facial motion by warps the photo in a warped photo attack. 2) Cut Photo Attack: The high resolution printed photos are then used for the cut photo attacks. In this scenario, an attacker hides behinds the photo

Fig. 3. Sample from the CASIA FASD. From top to bottom: low, normal and high quality images. From the left to the right: real faces and warped photo, cut photo and video replay attacks.

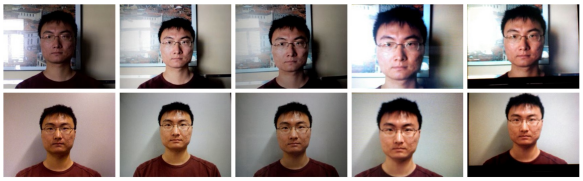


Fig. 4. Samples from the REPLAY-ATTACK database. The first row presents images taken from the controlled scenario, while the second row corresponds to the images from the adverse scenario. From the left to the right: real faces and high definition, mobile and print attacks.

Fig. 5. Samples from the OULU-NPU database. From top to bottom is the three sessions with different acquisition conditions. From the left to the right: real faces, print attack 1, print attack 2, video attack 1 and video attack 2.

and exhibits eye-blinking through the holes of the eye region, which was cut off before attack. In addition, the attacker put a intact photo behind the cut photo, putting the eye region overlapping from the holes and moving the intact photo up and down slightly to simulate the blinking of the eyes. 3) Video Attack: In this attack, the high resolution videos are displayed on an iPad and captured by a camera.

2) REPLAY-ATTACK Database: The REPLAY-ATTACK Database consists of video recordings of real accesses and attack attempts to 50 clients (see, Fig.4). There are 1200 videos taken by the webcam on a MacBook with the resolution 320 $\times$ 240 under two illumination conditions: 1) controlled condition with a uniform background and light supplied by a fluorescent lamp, 2) adverse condition with non-uniform background and the day-light. For performance evaluation, the data set is divided into three subsets of training (360 videos), development (360 videos), and testing (480 videos). To generate the fake faces, a high resolution videos were taken for each person using a Canon PowerShot camera and an iPhone 3GS camera, under the same illumination conditions. Three types of attacks were designed: (1) Print Attacks: High resolution pictures were printed on A4 paper and recaptured by cameras; (2) Mobile Attacks: High resolution pictures and videos were displayed on the screen of an iPhone 3GS and recaptured by cameras; (3) High Definition Attacks: the pictures and the videos were displayed on the screen of an iPad with resolution of 1024 $\times$ 168.

3) OULU-NPU Database: OULU-NPU face presentation attack database consists of 4950 real access and attack videos that were recorded using front facing cameras of six different mobile phones (see, Fig.5). The real videos and attack materials were collected in three sessions with different illumination condition. The attack types considered in the OULU-NPU database are print and video-replay. These attacks were created using two printers (Printer 1 and 2) and two display devices (Display 1 and 2). The videos of the real accesses and attacks, corresponding to the 55 subjects, are divided into three subject-disjoint subsets for training, development and testing with 20, 15 and 20 users, respectively.

B. Experimental Settings

In our experiments, we followed the protocols associated with each of the three databases which allows a fair comparison with other methods proposed in the state of art. For CASIA FASD, the model parameters are trained and tuned using the training set and the results are reported in terms of Equal Error Rate (EER) on the test set. Since the REPLAY-ATTACK database provides a validation set, the results are given in terms of EER on the validation set and the Half Total Error Rate (HTER) on the test set following the official test protocol. EER is achieved at the point where the false rejection rate (FRR) is equal to false acceptance rate (FAR). To compute HTER, we first compute EER and the corresponding threshold on the validation set. Then HTER can be calculated via the threshold on the test set.

Following [58], we evaluate our method on OULU-NPU database with two metrics: Attack Presentation Classification Error Rate (APCER) (Eq. (14)) and Bona Fide Presentation Classification Error Rate (BPCER) (Eq. (15)).

$$APCER_{PAI} = \frac{1}{N_{PAI}} \sum_{i=1}^{N_{PAI}} (1 - R_{es_i}) \qquad (14)$$

$$BPCER = \frac{\sum_{i=1}^{N_{BF}} R_{es_i}}{N_{BF}} \quad (15)$$

where, N_{PAI} is the number of the attack presentations for the certain presentation attack instruments (PAI), N_{BF} is the total number of the bona fide presentations. If the prediction of ith presentation is attack, R_{es_i} gets the value "1", while the prediction is bona fide, the value of R_{es_i} is "0". These two metrics correspond to the False Acceptance Rate (FAR) and False Rejection Rate (FRR) commonly used in the PAD related literature [58], [59]. In addition, we apply the average of the APCER and the BPCER, called Average Classification Error Rate (ACER), to measure the overall performances.

For the operational systems, the metrics we used (EER, HTER, APCER and BPCER) cannot quantify verification performance. Following the Face Recognition Vendor Test (FRVT) and the common metrics of face recognition, the Receiver Operating Characteristic (ROC) is used to measure the performance of liveness detection. To clearly visualize the TPR@FAR=0.1 and TPR@FAR=0.01 in the figures, the logarithmic coordinates are used for the X-axis of the ROC curves.

To be consistent with many previous works, the pre-processing steps are needed, consisted of frame sampling and face alignment. Since these three databases consist of videos, we extract the frames from each video. After that, the MTCNN [51] is used for face detection and landmark detection. Then the detected faces are aligned to size of 128 $\times$ 128. For every aligned face, we conduct data augmentation including horizontal flipping, random rotation (0-20 degree), and random crop (114 $\times$ 114).

For each database, we used the training set to fine-tune the MobileNet and ResNet-18 model with cross-entropy loss and the testing set and validation set are used to evaluate the performance.

For the learning parameter setting, we set the momentum as 0.9 and the learning rate as 0.0001 for training the network. It is observed that the network training converges after 50 epochs with the batch size 128 during the training.

C. Results of CASIA-FASD

The CASIA-FASD is split into the training set comprised of 20 subjects and the test set containing 30 individuals. For each of the seven attacking scenarios, the data should then be selected from the corresponding training and test sets for model training and evaluation.

Different color spaces might lead to different performance of anti-spoofing [48], though RGB color is the most widely used. To explore the effect of color space, we conduct experiments and compare the performance of three color spaces: RGB, HSV and YCbCr. All the training settings of 3 color space keep the same. Specifically, the original input images/frames in database are converted to MSR images. Then the images of different color spaces are fed to our TSCNN respectively. The spoofing detection results (EER, the lower the better) based on MobileNet and ResNet-18 are reported in Table I. The ROC curves are shown in Fig.6-(a) and the attention Fusion results in terms of TPR@FAR=0.1 and TPR@FAR=0.01 are presented in Table VII.

Results: (1) From results on seven scenarios, RGB and YCbCr generally outperform HSV color space using both ResNet-18 and MobileNet. And the results of RGB and YCbCr are quite similar.

(2) We can see that RGB, HSV and YCbCr features all work better than MSR features for both MobileNet (4.931%, 5.134% and 5.091% vs. 9.531%) and ResNet-18 (3.437%, 4.831% and 3.635 vs. 7.883%).

(3) The fusion of MSR and RGB features works better than MSR and HSV, MSR and YCbCr for both MobileNet (4.175% VS 5.061% and 4.339%) and ResNet-18 (3.145% VS 4.661% and 4.761%). (4) The fusion of MSR and RGB features works better than individual one for MobileNet (fusion: 4.175% vs RGB: 4.931% and MSR: 9.513%). The same conclusion can be drawn for ResNet-18 fusion. As for the reason why RGB is better than HSV and YCbCr, we believe that the MSR plays a role of reducing the impact of illuminations, while the RGB tries to preserve the detailed facial textures. However, HSV and YCbCr are based on the separation of the luminance and the chrominance, which are not effective for the fusion with MSR. It verifies the complementarity of RGB and MSR images.

(4) From the Table VII, not surprisingly, the overall results of CASIA-FASD with ResNet (99.71% and 85.33%) are better than that with MobileNet (98.95% and 82.51%).

D. Results of REPLAY-ATTACK and OULU-NPU

REPLAY-ATTACK and OULU-NPU are divided into three subsets: training, test and development. The training set is used to train a classifier or feature extractor while the development set is typically employed to adjust parameters of the classifier. The test set is used for result evaluation. In this experiment, we follow the experimental settings of CASIA-FASD and use MobileNet and ResNet-18 for evaluation.

From Table II and Fig.6-b, we can see the fusion of MSR and RGB works better than individual ones in terms of EER (fusion: 0.131% vs RGB: 0.384% and MSR: 7.365%) and HTER (fusion: 0.254% vs RGB: 1.561% and MSR: 8.584%) on REPLAY-ATTACK database using MobileNet. The same conclusion can be found for ResNet-18. From Table VII, the overall results of REPLAY-ATTACK using MobileNet (99.42% and 99.13%) are better than that with ResNet-18 (99.21% and 98.59%). In addition, we further fuse the fused MobileNet features (RGB+MSR) and fused ResNet-18 features (RGB+MSR). Because feature dimensionality of original MobileNet (1024D) and ResNet-18 (512D) is different, we change the bottleneck layer of the MobileNet to be of 512D to conduct our attention-based fusion. From Table II, we can see this further fusion works better than ResNet fusion, but slightly worse than the MobileNet fusion.

To further verify the effectiveness of the fusion of RGB and MSR on illumination variations, we conduct the experiment on REPLAY-ATTACK database which contains two illumination conditions: 1) controlled condition with a uniform background and light supplied by a fluorescent lamp, 2) adverse condition with non-uniform background and the day-light. To discuss

TABLE I
EER (%) OF THREE COLOR SPACES AND MSR FEATURES ON CASIA-FASD DATABASE IN SEVEN SCENARIOS

	Attack Scenarios	Low	Normal	High	Warped	Cut	Video	Overall
LBP	RGB	15.301	8.996	6.412	8.551	6.011	5.661	7.802
	MSR	10.690	10.302	5.331	7.609	8.091	8.701	9.003
	RGB+MSR Fusion	8.996	9.330	5.981	7.604	6.771	4.390	7.408
MobileNet	RGB	10.610	4.606	5.260	5.934	3.978	3.846	4.931
	HSV	8.714	5.884	6.995	3.723	4.709	4.682	5.143
	YCbCr	8.441	4.993	4.519	6.410	5.792	3.904	5.091
	MSR	7.056	8.129	5.818	9.828	5.126	9.833	9.531
	RGB+MSR Fusion	6.745	4.068	3.258	5.258	2.453	**2.647**	4.175
	HSV+MSR Fusion	7.633	4.982	5.601	4.679	4.510	4.511	5.061
	YCbCr+MSR Fusion	7.003	5.120	3.227	4.031	6.001	3.799	4.339
ResNet	RGB	4.021	5.851	**1.703**	5.019	**1.941**	2.679	3.437
	HSV	6.341	2.291	5.815	3.459	2.992	4.578	4.831
	YCbCr	7.441	2.185	1.713	4.249	3.329	3.716	3.635
	MSR	6.793	6.270	10.098	7.665	5.087	9.531	7.883
	RGB+MSR Fusion	**3.545**	**2.170**	2.785	4.419	2.572	4.931	**3.145**
	HSV+MSR Fusion	5.319	2.907	4.886	**3.299**	2.555	4.931	4.661
	YCbCr+MSR Fusion	6.178	3.099	4.690	4.003	3.133	3.999	4.761

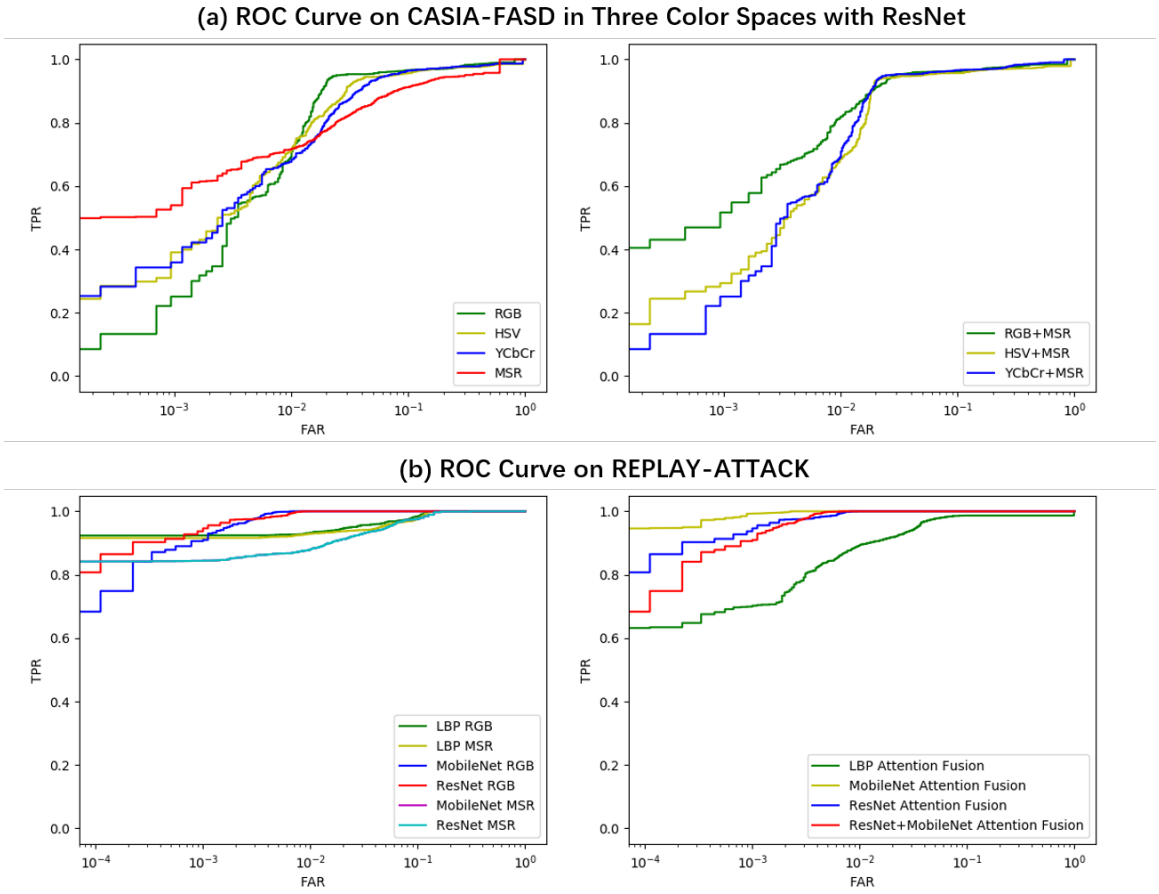


Fig. 6. ROC curves on REPLAY-ATTACK and CASIA-FASD databases. (a) ROC curves on CASIA-FASD with ResNet under different color spaces and MSR. (b) ROC curves on REPLAY-ATTACK with MobileNet with LBP and CNNs.

the improvements over lightings, we divided the database into two parts: adverse illumination and controlled illumination and run the experiments separately. From Table III and Fig.7-(a), MSR features have the better results than RGB features in adverse illumination (stronger lighting), showing the robustness of MSR on strong lightings. On the other hand, RGB outperforms MSR features in controlled illumination (close to neutral lighting), showing the RGB has the strong capacity to maintain the texture details under neutral illuminations. After fusion, the results are improved in both adverse and controlled illumination. So the Fusion of MSR and RGB can effectively handle various lightings and improve the performance.

For the OULU-NPU database, we follow [58] to use four metrics: we present EER in development set and APCER, BPCER and ACER in test set.

Table IV and Table VII shows the results of RGB, MSR

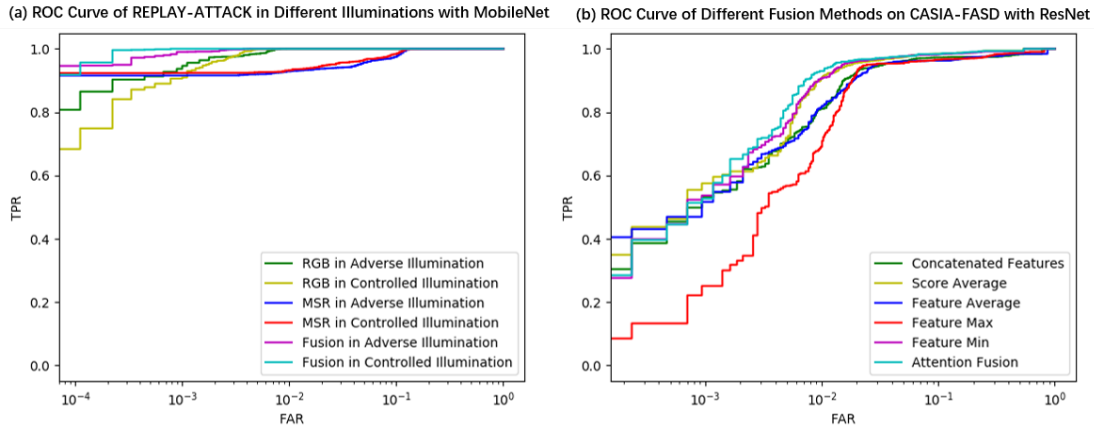


Fig. 7. ROC curves on on REPLAY-ATTACK and CASIA-FASD databases. (a) ROC curves on REPLAY-ATTACK with MobileNet under different illuminations. (b) ROC curves on CASIA-FASD with ResNet under different fusion methods.

TABLE II
EER (%) AND HTER (%) OF RGB AND MSR FEATURES ON REPLAY-ATTACK DATABASE

Method	REPLAY-ATTACK	
	EER	HTER
LBP RGB	3.990	4.788
LBP MSR	4.701	5.060
MobileNet RGB	0.384	1.561
ResNet RGB	0.628	2.038
MobileNet MSR	7.365	8.584
ResNet MSR	8.350	9.576
LBP Attention Fusion	3.491	4.903
MobileNet Attention Fusion	**0.131**	**0.254**
ResNet Attention Fusion	0.210	0.389
ResNet + MobileNet Attention Fusion	0.177	0.293

TABLE III
EER (%) AND HTER (%) OF RGB AND MSR FEATURES ON ADVERSE ILLUMINATION AND CONTROLLED ILLUMINATION IN REPLAY-ATTACK DATABASE

Method	Adverse illumination		Controlled illumination	
	EER	HTER	EER	HTER
MobileNet RGB	0.451	1.971	0.140	1.107
ResNet RGB	0.705	2.444	0.411	1.677
MobileNet MSR	7.660	8.621	6.138	7.218
ResNet MSR	8.720	9.031	7.993	8.930
MobileNet Attention Fusion	**0.165**	**1.299**	**0.093**	**0.097**
ResNet Attention Fusion	0.285	1.433	0.169	1.310

and fusion feature based on MobileNet and ResNet-18. In terms of ACER and EER, we can see the fusion of RGB and MSR performs better than individual ones. For most results in four protocols, the fusion of features significantly outperforms individual features.

The consistent improvement of feature fusion shows the effectiveness of the use of two information sources: RGB and MSR. As shown in Table II and Table IV, the popular networks (MobileNet and ResNet-18) achieve competitive performances on REPLAY-ATTACK and OULU-NPU database .

TABLE IV
EER (%), APCER (%) , BPCER (%) AND ACER (%) OF RGB AND MSR FEATURES ON OULU-NPU DATABASE

Prot.	Methods	Dev	Test		
		EER(%)	APCER(%)	BPCER(%)	ACER(%)
1	MobileNet RGB	6.1	9.6	**6.2**	7.9
	ResNet RGB	2.3	**3.5**	8.7	6.1
	MobileNet MSR	10.5	10.6	9.4	10.0
	ResNet MSR	5.7	7.5	9.3	8.4
	MobileNet Attention Fusion	5.2	3.9	9.5	6.7
	ResNet Attention Fusion	**2.1**	5.1	6.7	**5.9**
2	MobileNet RGB	5.7	6.5	10.7	8.6
	ResNet RGB	2.7	3.7	8.1	5.9
	MobileNet MSR	9.6	8.9	9.9	9.4
	ResNet MSR	4.3	3.8	11.6	7.8
	MobileNet Attention Fusion	5.1	**3.6**	9.0	6.3
	ResNet Attention Fusion	**2.0**	7.6	**2.2**	**4.9**
3	MobileNet RGB	5.3±0.5	**3.5±1.8**	9.3±2.6	6.4±3.7
	ResNet RGB	2.7±0.8	9.3±0.8	5.7±1.2	7.2±2.6
	MobileNet MSR	10.8±1.2	6.9±2.5	12.3±0.9	9.7±1.9
	ResNet MSR	4.6±0.8	8.3±1.9	9.4±1.8	8.7±2.1
	MobileNet Attention Fusion	5.1±0.3	8.7±4.5	**5.3±2.3**	6.3±2.2
	ResNet Attention Fusion	**1.9±0.4**	3.9±2.8	7.3±1.1	**5.6±1.6**
4	MobileNet RGB	6.3±0.4	12.3±7.5	9.7±2.6	10.3±3.1
	ResNet RGB	2.6±0.5	17.9±9.1	10.1±5.5	14.9±6.4
	MobileNet MSR	11.8±1.8	24.7±10.5	21.3±12.8	22.0±11.6
	ResNet MSR	6.6±0.7	19.6±9.1	16.2±8.8	17.1±8.1
	MobileNet Attention Fusion	6.1±0.7	**10.9±4.6**	12.7±5.1	11.3±3.9
	ResNet Attention Fusion	**2.3±0.3**	11.3±3.9	**9.7±4.8**	**9.8±4.2**

E. Attention based fusion results

As mentioned above, RGB feature is mainly focusing on micro-texture of facial skin on the all frequencies together, while the MSR feature is focusing on the high frequencies which reduces the influence of illumination. Table I, Table II and Table IV have verified the effectiveness of the fusion of these two features (RGB and MSR). In this section, we further explore this effectiveness.

TABLE V
EER (%) OF DIFFERENT FUSION METHODS ON CASIA-FASD DATABASES IN SEVEN SCENARIOS

	Attack Scenarios	Low	Normal	High	Warped	Cut	Video	Overall
MobileNet	Concatenated Features	7.808	3.473	5.957	5.364	3.267	4.479	5.191
	Score Average	10.611	4.612	5.312	5.934	3.971	3.877	4.953
	Feature Average	8.086	**3.311**	5.819	5.253	3.333	4.278	5.108
	Feature Max	8.048	3.410	6.017	5.347	3.321	4.529	5.201
	Feature Min	7.820	3.458	5.380	**5.149**	3.267	4.064	4.887
	Attention Fusion	**6.745**	4.068	**3.258**	5.258	**2.453**	**2.647**	**4.175**
ResNet	Concatenated Features	5.568	3.099	4.302	**4.092**	**2.516**	**3.143**	3.380
	Score Average	5.902	2.969	3.830	4.202	2.658	3.224	3.332
	Feature Average	6.242	3.291	4.689	3.935	2.929	3.956	3.895
	Feature Max	5.846	4.039	4.536	4.331	3.091	4.198	4.189
	Feature Min	7.244	2.825	4.941	4.280	3.030	3.984	4.157
	Attention Fusion	**3.545**	**2.170**	**2.785**	4.419	2.572	4.931	**3.145**

First, we show some qualitative results via visualization. Compared with average feature fusion which weights different features equally, attention fusion has the flexibility to adaptively weight the features in an asymmetry way. Therefore, our attention-based fusion has the potential to obtain the better weights leading to better performance. Fig.8-(A) shows this asymmetry weighting mechanism of our attention-based fusion method. The samples in Fig.8-(A) are selected from REPLAY-ATTACK database which covers two imaging lightness conditions: adverse illumination (uneven, complicated lightings), controlled illumination (even, neutral lightings). From the samples in Fig.8, we can see the weights for MSR and RGB are adaptively asymmetry. Under adverse (uneven, complicated lightings) illumination, the weights of MSR images are higher than those of RGB ones because MSR images are more illumination-invariant than RGB ones. Under controlled illumination, unsurprisingly, the RGB images gain higher weights. Fig.8 (B) shows some samples under different illuminations with three scores (RGB, MSR, the fusion of them). We can see some samples failed with individual RGB or MSR scores, but the fusion results lead to correct recognition, showing the effectiveness of the fusion of RGB and MSR, in particular, under various illuminations.

Second, we show some qualitative results. Specifically, we compare the proposed attention-based fusion methods with some popular feature fusion methods including score averaging, feature concatenation, feature averaging, feature max pooling, feature min pooling and the proposed attention method. The fusion results are presented separately for different databases.

Table V shows the results of CASIA-FASD with the seven scenarios. In addition, Fig.7-(b) shows the ROC curves of the popular feature fusion methods using MobileNet. The proposed attention based fusion method achieves the lowest EER across all other scenarios ('Overall') 4.175% (MobileNet) and 3.145% (ResNet-18), showing that the superiority of the our fusion methods against others. For MoblieNet and ResNet-18, the 2nd and 3rd best performed fusion methods are {'Feature Min' and 'Score Average'} and {'Score Average' and 'Concatenated Features'}, respectively.

Table IV-E shows the fusion results on REPLAY-ATTACK and OULU-NPU. We can see that our attention-based fusion works consistently better than all other fusion methods on both REPLAY-ATTACK (EER and HTER) and OULU-NPU (EER). The promising performance results from the fact that attention-based fusion can adaptively weight the RGB and MSR features.

TABLE VI
EER (%) AND HTER (%)OF DIFFERENT FUSION METHODS ON REPLAY-ATTACK AND OULU-NPU DATABASES

	Methods	REPLAY-ATTACK		OULU-NPU
		EER	HTER	EER
MobileNet	Concatenated Features	0.412	0.381	6.381
	Score Average	0.363	0.360	6.472
	Feature Average	0.396	0.395	7.549
	Feature Max	0.310	0.294	8.317
	Feature Min	0.574	0.565	9.841
	Attention Fusion	**0.131**	**0.254**	**5.692**
ResNet	Concatenated Features	0.841	0.668	4.518
	Score Average	1.278	1.178	9.565
	Feature Average	0.873	0.725	5.358
	Feature Max	0.958	0.906	4.964
	Feature Min	0.579	0.490	2.578
	Attention Fusion	**0.210**	**0.389**	**2.021**

F. Comparisons with State-of-the-art

Table VIII presents the comparisons of our approach with the state-of-the-art methods for face spoofing detection. In general, the proposed algorithm outperforms many competitors, demonstrating the effectiveness of our method by fusing RGB feature and MSR feature with attention model.

For REPLAY-ATTACK database, the proposed method achieves the best (MobileNet+Attention) and 2nd best (ResNet-18+Attention) performance in terms of EER, showing the effectiveness of the fusion of two clues (RGB and MSR). In terms of HTER, our method (MobileNet+Attention) achieves the 2nd best performance, slightly lower than Bottleneck feature fusion + NN [50]. However, our method greatly outperforms [50] in terms of EER.

For CASIA-FASD database, it can be seen in Table VIII that we also achieve the best (ResNet-18 + Attention) and 2nd best (MobileNet + Attention) performance in terms of EER.

For OULU-NPU database, as shown in Table IX, we can achieve 2nd best performance for most results under the four protocols, while the method of [63] works best, which uses the additional information of 3D depth shape and rPPG (The rPPG signal provides temporal information about face liveness,

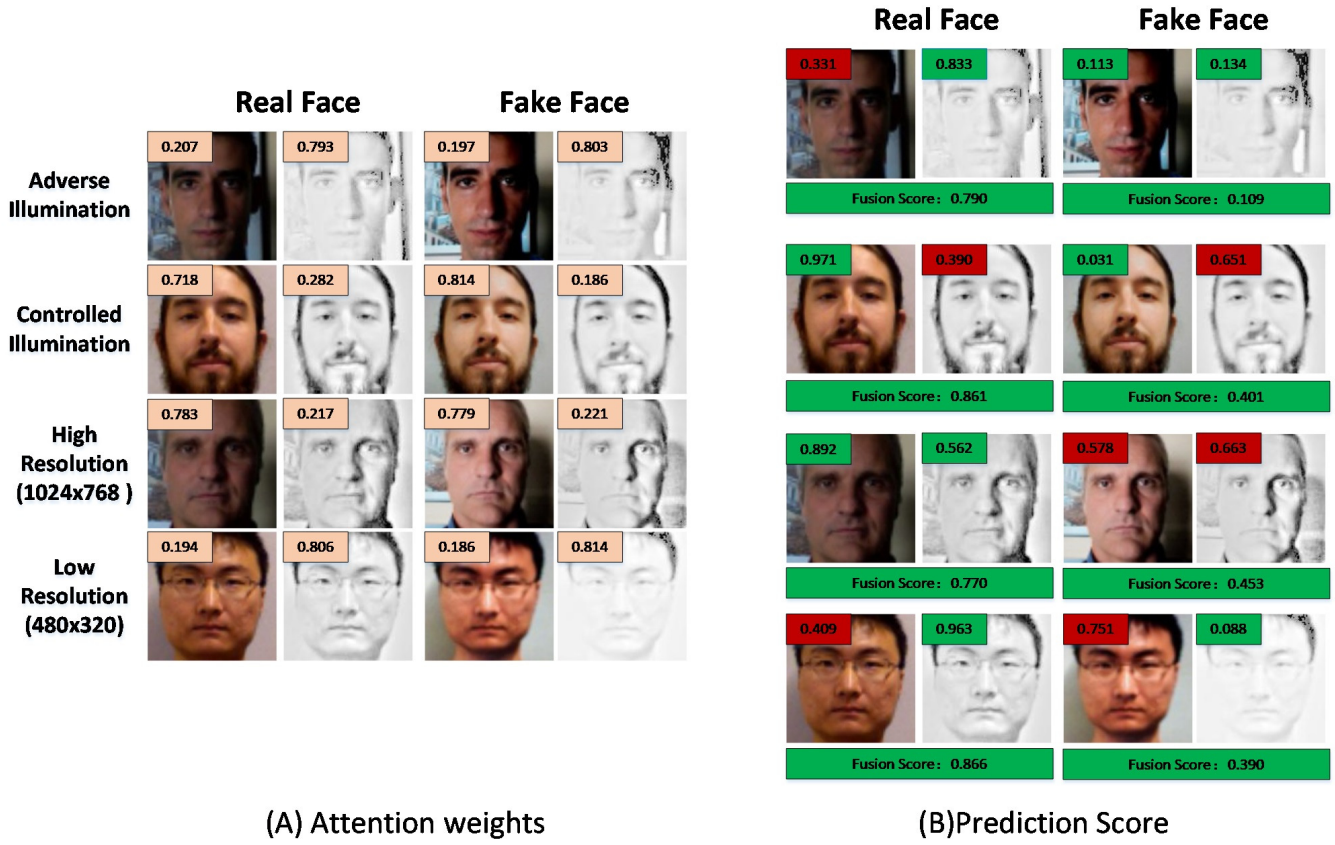


Fig. 8. Results on REPLAY-ATTACK database. (A) Attention fusion weights (numbers in the boxes) showing the importance of RGB and MSR. Samples cover 2 imaging lightness conditions: adverse illumination (Row 1 and 2) and controlled illumination (Row 3 and 4). (B) Three prediction scores: RGB, MSR and the fusion of them (numbers in the boxes). The red and green boxes indicate the wrong and correct predictions respectively.

TABLE VII
TPR@FAR=0.1 AND TPR@FAR=0.01 OF THE ATTENTION FUSION RESULTS ON CASIA-FASD, REPLAY-ATTACK AND OULU-NPU DATABASES

Database	Methods	Protocol	TPR@FAR=0.1	TPR@FAR=0.01
CASIA-FASD	ResNet Attention Fusion	overall	99.71%	85.33%
	MobileNet Attention Fusion	overall	98.95%	82.51%
REPLAY-ATTACK	ResNet Attention Fusion	overall	99.21%	98.59%
	MobileNet Attention Fusion	overall	99.42%	99.13%
OULU-NPU	ResNet Attention Fusion	Prot.1	94.15%	83.44%
		Prot.2	95.11%	86.78%
		Prot.3	93.59%±0.5%	84.39%±0.4%
		Prot.4	93.09%±0.4%	83.69%±0.5%
	MobileNet Attention Fusion	Prot.1	98.94%	96.74%
		Prot.2	99.10%	96.86%
		Prot.3	98.41%±0.6%	96.04%±0.5%
		Prot.4	97.83%±0.4%	95.22%±0.6%

which is related to the intensity changes of facial skin over time).

To summarize, our method can achieve very strong performance across all the three benchmark databases, showing the merits of the proposed method.

G. Cross-Database Comparisons

The spoofing faces of different databases are captured using different devices under different environments (e.g. lightings). Therefore, it is interesting to evaluate our strategy in a cross-database protocol to verify its generalization capacity.We conducted a cross-database evaluation between CASIA-FASD and REPLAY-ATTACK. To be more specific, cross-database is to train and tune the classifier on one database and test on another database. The generalization ability of the system in this case is manifested by the HTER obtained on the validation and test sets. The countermeasure was trained and tuned with CASIA-FASD or REPLAY-ATTACK each time, and then tested on the other databases. The results are reported in Table IV-F compared with the state-of-the-art techniques in this cross-database manner.

Due to the domain shift (different imaging environments) between databases, the performaence of all the anti-spoofing methods drops. Compared with the state-of-the-art methods,

TABLE VIII
COMPARISON BETWEEN THE PROPOSED COUNTERMEASURE AND STATE-OF-THE-ART METHODS ON REPLAY-ATTACK AND CASIA-FASD DATABASES IN TERMS OF EER(%) AND HTER(%)

Methods	REPLAY-ATTACK		CASIA-FASD
	EER	HTER	EER
Motion [60]	11.6	11.7	26.6
LBP [56]	13.9	13.8	18.2
LBP-TOP [61]	7.90	7.60	10.00
CDD [62]	-	-	11.8
DOG [3]	-	-	17.0
DMD [27]	5.3	3.8	21.8
IQA [4]	-	15.2	32.4
CNN [14]	6.10	2.10	7.40
IDA [5]	-	7.4	-
Motion + LBP [29]	4.50	5.11	-
Color-LBP [10]	0.40	2.90	6.20
Bottleneck feature fusion + NN [50]	0.83	**0.00**	5.83
Ours (MobileNet + Attention)	**0.131**	0.254	4.175
Ours (ResNet-18 + Attention)	0.210	0.389	**3.145**

TABLE IX
COMPARISON BETWEEN THE PROPOSED COUNTERMEASURE AND STATE-OF-THE-ART METHODS ON OULU-NPU DATABASE IN TERMS OF EER (%), APCER (%), BPCER (%) AND ACER (%)

Prot.	Methods	Dev		Test	
		EER(%)	APCER(%)	BPCER(%)	ACER(%)
1	CpqD [58]	0.6	2.9	10.8	6.9
	GRADANT [58]	1.1	1.3	12.5	6.9
	Depth + rPPG [63]	-	1.6	1.6	1.6
	MobileNet Attention Fusion	5.2	3.9	9.5	6.7
	ResNet Attention Fusion	2.1	5.1	6.7	5.9
2	MixedFASNet [58]	1.3	9.7	2.5	6.1
	GRADANT [58]	0.9	3.1	1.9	2.5
	Depth + rPPG [63]	-	2.7	2.7	2.7
	MobileNet Attention Fusion	5.1	3.6	9.0	6.3
	ResNet Attention Fusion	2.0	7.6	2.2	4.9
3	MixedFASNet [58]	1.4±0.5	5.3±6.7	7.8±5.5	6.5±4.6
	GRADANT [58]	0.9±0.4	2.6±3.9	5.0±5.3	3.8±2.4
	Depth + rPPG [63]	-	2.7±1.3	3.1±1.7	2.9±1.5
	MobileNet Attention Fusion	5.1±0.3	8.7±4.5	5.3±2.3	6.3±2.2
	ResNet Attention Fusion	1.9±0.4	3.9±2.8	7.3±1.1	5.6±1.6
4	Massy HNU [58]	1.0±0.4	35.8±35.3	8.3±4.1	22.1±17.6
	GRADANT [58]	1.1±0.3	5.0±4.5	15.0±7.1	10.0±5.0
	Depth + rPPG [63]	-	9.3±5.6	10.4±6.0	9.5±6.0
	MobileNet Attention Fusion	6.1±0.7	10.9±4.6	12.7±5.1	11.3±3.9
	ResNet Attention Fusion	2.3±0.3	11.3±3.9	9.7±4.8	9.8±4.2

TABLE X
INTER-DATABASE TEST RESULTS IN TERMS OF HTER (%) ON THE CASIA-FASD AND REPLAY-ATTACK DATABASE

Method	Train	Test	Train	Test
	CASIA FASD	REPLAY ATTACK	REPLAY ATTACK	CASIA FASD
Motion [60]	50.2%		47.9%	
LBP [56]	55.9%		57.6%	
LBP-TOP [61]	49.7%		60.6%	
Motion-Mag [64]	50.1%		47.0%	
Spectral cubes [22]	34.4%		45.5%	
CNN [14]	48.5%		39.6%	
Color-LBP [10]	47.0%		39.6%	
Colour Texture [8]	30.3%		37.7%	
Depth + rPPG [63]	27.6%		28.4%	
Deep-Learning [13]	48.2%		45.4%	
KSA [65]	33.1%		32.1%	
Frame difference [66]	50.25%		43.05%	
Ours (MobileNet + Attention)	30.0%		33.4%	
Ours (ResNet-18 + Attention)	36.2%		34.7%	

TABLE XI
INTER-DATABASE TEST RESULTS FOR RGB FEATURES IN TERMS OF MAXIMUM MEAN DISCREPANCY ON THE CASIA-FASD AND REPLAY-ATTACK DATABASE

Model	Train	Val	MMD
Resnet18 RGB	CASIA-FASD	CASIA-FASD	0.7653
	CASIA-FASD	REPLAY-ATTACK	1.4561
	REPLAY-ATTACK	REPLAY-ATTACK	**0.6871**
	REPLAY-ATTACK	CASIA-FASD	1.3484
Mobilenet RGB	CASIA-FASD	CASIA-FASD	0.8654
	CASIA-FASD	REPLAY-ATTACK	1.3276
	REPLAY-ATTACK	REPLAY-ATTACK	0.7469
	REPLAY-ATTACK	CASIA-FASD	1.2765

TABLE XII
INTER-DATABASE TEST RESULTS FOR MSR FEATURES IN TERMS OF MAXIMUM MEAN DISCREPANCY ON THE CASIA-FASD AND REPLAY-ATTACK DATABASE

Model	Train	Val	MMD
Resnet18 MSR	CASIA-FASD	CASIA-FASD	0.9831
	CASIA-FASD	REPLAY-ATTACK	1.8746
	REPLAY-ATTACK	REPLAY-ATTACK	**0.6541**
	REPLAY-ATTACK	CASIA-FASD	1.0133
Mobilenet MSR	CASIA-FASD	CASIA-FASD	0.8655
	CASIA-FASD	REPLAY-ATTACK	1.7749
	REPLAY-ATTACK	REPLAY-ATTACK	0.8811
	REPLAY-ATTACK	CASIA-FASD	1.1661

TABLE XIII
INTER-DATABASE TEST RESULTS FOR RGB AND MSR FUSION FEATURES IN TERMS OF MAXIMUM MEAN DISCREPANCY ON THE CASIA-FASD AND REPLAY-ATTACK DATABASE

Model	Train	Val	MMD
Resnet18 RGB + MSR Fusion	CASIA-FASD	CASIA-FASD	**0.6215**
	CASIA-FASD	REPLAY-ATTACK	1.2511
	REPLAY-ATTACK	REPLAY-ATTACK	0.7003
	REPLAY-ATTACK	CASIA-FASD	1.1295
Mobilenet RGB + MSR Fusion	CASIA-FASD	CASIA-FASD	0.6619
	CASIA-FASD	REPLAY-ATTACK	1.3518
	REPLAY-ATTACK	REPLAY-ATTACK	0.7139
	REPLAY-ATTACK	CASIA-FASD	1.0551

our method (MobileNet + Attention) achieves the 2nd best performance (30.0% and 33.4%), slightly worse than the best one [63] (27.6% and 28.4%). However, [63] uses more auxiliary information (3D face shape, rPPG signals) than our method.

To explore the reasons of performance drop in the cross-database evaluation, we consider the standard distribution distance metric, maximum mean discrepancy (MMD) [67] to measure the distance domain shift between the source feature and target feature distributions.

$$MMD(F_T, F_V) = \left\| \frac{1}{|F_T|} \sum_{f_t \in F_T} \phi(f_t) - \frac{1}{|F_V|} \sum_{f_v \in F_V} \phi(f_v) \right\| \quad (16)$$

As shown in the equation above, we define a representation $\phi()$, which operates on train data features, $f_t \in F_T$ and validate data features, $f_v \in F_V$. The larger the value of MMD, the bigger the domain shift.

From the result of Table XI XII XIII, we can see that: (1) When we train and test on the same database, the MMD is smaller than that train and test on different databases for both MobileNet and ResNet-18.

(2) Since the CASIA-FASD has seven scenarios, when we train on the CASIA-FASD database and test on the REPLAY-ATTACK database, the MMD is bigger than that we train on the REPLAY-ATTACK and test on the CASIA-FASD database.

(3) The fusion of RGB and MSR features reduced the MMD of the cross-database compared with individual one for both MobileNet and ResNet-18.

V. Conclusion

In this paper, we proposed an attention-based two stream convolutional networks for face spoofing detection to distinguish real and fake faces. The proposed approach applies the complementary features (RGB and MSR) extracted via CNN models (MobileNet and ResNet-18) and then employs the attention based fusion method to fuse these two features. The adaptively weighted features contain more discriminative information under various lighting conditions.

We evaluated our approaches of face spoofing on three challenging databases, i.e. CASIA-FASD, REPLAY-ATTACK and OULU-NPU, which indicated the competitive performance in both intra-database and inter-database. The experiments of fusion methods show that the attention model can achieve promising results on feature fusion. The cross-database evaluations show the effectiveness of the fusion of RGB and MSR information.

Acknowledgement

The authors would like to thank the journal reviewers for their valuable suggestions. This work was supported in part by the National Natural Science Foundation of China (61876072, 61876178, 61872367, 61572501) and the Fundamental Research Funds for the Central Universities.

References

[1] J. Li, Y. Wang, and A. K. Jain, "Live face detection based on the analysis of fourier spectra," *Proc Spie*, vol. 5404, pp. 296–303, 2004.

[2] 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*, 2010, pp. 504–517.

[3] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face antispoofing database with diverse attacks," in *Iapr International Conference on Biometrics*, 2012, pp. 26–31.

[4] J. Galbally and S. Marcel, "Face anti-spoofing based on general image quality assessment," in *22nd International Conference on Pattern Recognition, ICPR 2014, Stockholm, Sweden, August 24-28, 2014*, 2014, pp. 1173–1178.

[5] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," *IEEE Trans. Information Forensics and Security*, vol. 10, no. 4, pp. 746–761, 2015.

[6] J. Yan, Z. Zhang, Z. Lei, D. Yi, and S. Z. Li, "Face liveness detection by exploring multiple scenic clues," in *12th International Conference on Control Automation Robotics & Vision, ICARCV 2012, Guangzhou, China, December 5-7, 2012*, 2012, pp. 188–193.

[7] K. Patel, H. Han, A. K. Jain, and G. Ott, "Live face video vs. spoof face video: Use of moiré patterns to detect replay video attacks," in *International Conference on Biometrics, ICB 2015, Phuket, Thailand, 19-22 May, 2015*, 2015, pp. 98–105.

[8] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face spoofing detection using colour texture analysis," *IEEE Trans. Information Forensics and Security*, vol. 11, no. 8, pp. 1818–1830, 2016.

[9] Z. Boulkenafet, J. Komulainen, X. Feng, and A. Hadid, "Scale space texture analysis for face anti-spoofing," in *International Conference on Biometrics, ICB 2016, Halmstad, Sweden, June 13-16, 2016*, 2016, pp. 1–6.

[10] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face anti-spoofing based on color texture analysis," in *2015 IEEE International Conference on Image Processing, ICIP 2015, Quebec City, QC, Canada, September 27-30, 2015*, 2015, pp. 2636–2640.

[11] 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.*, vol. 24, no. 7, pp. 971–987, 2002.

[12] J. Määttä, A. Hadid, and M. Pietikäinen, "Face spoofing detection from single images using micro-texture analysis," in *2011 IEEE International Joint Conference on Biometrics, IJCB 2011, Washington, DC, USA, October 11-13, 2011*, 2011, pp. 1–7.

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

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

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *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, pp. 1106–1114.

[16] 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*, 2007, pp. 1–8.

[17] G. Pan, L. Sun, Z. Wu, and Y. Wang, "Monocular camera-based face liveness detection by combining eyeblink and scene context," *Telecommunication Systems*, vol. 47, no. 3-4, pp. 215–225, 2011.

[18] L. Sun, G. Pan, Z. Wu, and S. Lao, "Blinking-based live face detection using conditional random fields," in *Advances in Biometrics, International Conference, ICB 2007, Seoul, Korea, August 27-29, 2007, Proceedings*, 2007, pp. 252–260.

[19] A. Anjos, M. M. Chakka, and S. Marcel, "Motion-based countermeasures to photo attacks in face recognition," *Iet Biometrics*, vol. 3, no. 3, pp. 147–158, 2014.

[20] K. Kollreider, H. Fronthaler, M. I. Faraj, and J. Bigun, "Real-time face detection and motion analysis with application in "liveness" assessment," *IEEE Transactions on Information Forensics Security*, vol. 2, no. 3, pp. 548–558, 2015.

[21] Y. Kim, J. Na, S. Yoon, and J. Yi, "Masked fake face detection using radiance measurements," *J Opt Soc Am A Opt Image Sci Vis*, vol. 26, no. 4, pp. 760–766, 2009.

[22] A. da Silva Pinto, H. Pedrini, W. R. Schwartz, and A. Rocha, "Face spoofing detection through visual codebooks of spectral temporal cubes," *IEEE Trans. Image Processing*, vol. 24, no. 12, pp. 4726–4740, 2015.

[23] D. J. Jobson, Z. Rahman, and G. A. Woodell, "A multiscale retinex for bridging the gap between color images and the human observation of scenes," *IEEE Trans. Image Processing*, vol. 6, no. 7, pp. 965–976, 1997.

[24] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua, "Neural aggregation network for video face recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 5216–5225.

[25] 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*, 2017, pp. 319–328.

[26] 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*, 2015, pp. 141–145.

[27] S. Tirunagari, N. Poh, D. Windridge, A. Iorliam, N. Suki, and A. T. Ho, "Detection of face spoofing using visual dynamics," *IEEE transactions on information forensics and security*, vol. 10, no. 4, pp. 762–777, 2015.

[28] Z. Zhang, D. Yi, Z. Lei, and S. Z. Li, "Face liveness detection by learning multispectral reflectance distributions," in *Ninth IEEE International Conference on Automatic Face and Gesture Recognition (FG 2011), Santa Barbara, CA, USA, 21-25 March 2011*, 2011, pp. 436–441. [Online]. Available: https://doi.org/10.1109/FG.2011.5771438

[29] J. Komulainen, A. Hadid, M. Pietikäinen, A. Anjos, and S. Marcel, "Complementary countermeasures for detecting scenic face spoofing attacks," in *International Conference on Biometrics, ICB 2013, 4-7 June, 2013, Madrid, Spain*, 2013, pp. 1–7.

[30] M. De Marsico, M. Nappi, D. Riccio, and J. Dugelay, "Moving face spoofing detection via 3d projective invariants," in *5th IAPR International Conference on Biometrics, ICB 2012, New Delhi, India, March 29 - April 1, 2012*, 2012, pp. 73–78.

[31] T. Wang, J. Yang, Z. Lei, S. Liao, and S. Z. Li, "Face liveness detection using 3d structure recovered from a single camera," in *International Conference on Biometrics, ICB 2013, 4-7 June, 2013, Madrid, Spain*, 2013, pp. 1–6.

[32] Y. Wang, F. Nian, T. Li, Z. Meng, and K. Wang, "Robust face anti-spoofing with depth information," *J. Visual Communication and Image Representation*, vol. 49, pp. 332–337, 2017.

[33] E. H. Land and J. J. Mccann, "Lightness and retinex theory," *J Opt Soc Am*, vol. 61, no. 1, pp. 1–11, 1971.

[34] D. Choi, I. H. Jang, M. H. Kim, and N. C. Kim, "Color image enhancement based on single-scale retinex with a jnd-based nonlinear filter," in *International Symposium on Circuits and Systems (ISCAS 2007), 27-20 May 2007, New Orleans, Louisiana, USA*, 2007, pp. 3948–3951.

[35] G. Zhang, D. Sun, P. Yan, H. Zhao, and Z. Li, "A LDCT image contrast enhancement algorithm based on single-scale retinex theory," in *2008 International Conferences on Computational Intelligence for Modelling, Control and Automation (CIMCA 2008), Intelligent Agents, Web Technologies and Internet Commerce (IAWTIC 2008), Innovation in Software Engineering (ISE 2008), 10-12 December 2008, Vienna, Austria*, 2008, pp. 181–186.

[36] D. J. Jobson, Z. Rahman, and G. A. Woodell, "Properties and performance of a center/surround retinex," *IEEE Trans. Image Processing*, vol. 6, no. 3, pp. 451–462, 1997.

[37] S. J. Xie, Y. Lu, S. Yoon, J. C. Yang, and D. S. Park, "Intensity variation normalization for finger vein recognition using guided filter based singe scale retinex," *Sensors*, vol. 15, no. 7, pp. 17 089–17 105, 2015.

[38] C. Lee, J. Shih, C. Lien, and C. Han, "Adaptive multiscale retinex for image contrast enhancement," in *Ninth International Conference on Signal-Image Technology & Internet-Based Systems, SITIS 2013, Kyoto, Japan, December 2-5, 2013*, 2013, pp. 43–50.

[39] E. Park, X. Han, T. L. Berg, and A. C. Berg, "Combining multiple sources of knowledge in deep cnns for action recognition," in *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016, Lake Placid, NY, USA, March 7-10, 2016*, 2016, pp. 1–8.

[40] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 2014, pp. 568–576.

[41] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[42] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan, "See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 6776–6785.

[43] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[44] R. Girdhar and D. Ramanan, "Attentional pooling for action recognition," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, 2017, pp. 33–44.

[45] A. Gupta, D. Agrawal, H. Chauhan, J. Dolz, and M. Pedersoli, "An attention model for group-level emotion recognition," *CoRR*, vol. abs/1807.03380, 2018.

[46] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 6450–6458.

[47] R. F. Nogueira, R. de Alencar Lotufo, and R. C. Machado, "Fingerprint liveness detection using convolutional neural networks," *IEEE Trans. Information Forensics and Security*, vol. 11, no. 6, pp. 1206–1213, 2016.

[48] L. Li, X. Feng, X. Jiang, Z. Xia, and A. Hadid, "Face anti-spoofing via deep local binary patterns," in *2017 IEEE International Conference on Image Processing, ICIP 2017, Beijing, China, September 17-20, 2017*, 2017, pp. 101–105.

[49] H. Li, P. He, S. Wang, A. Rocha, X. Jiang, and A. C. Kot, "Learning generalized deep feature representation for face anti-spoofing," *IEEE Trans. Information Forensics and Security*, vol. 13, no. 10, pp. 2639–2652, 2018.

[50] 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*, vol. 38, pp. 451–460, 2016.

[51] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multi-task cascaded convolutional networks," *CoRR*, vol. abs/1604.02878, 2016.

[52] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, 2017.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[54] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, 2009, pp. 248–255.

[55] 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*, 2012, pp. 26–31.

[56] 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*, 2012, pp. 1–7.

[57] 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*, 2017, pp. 612–618.

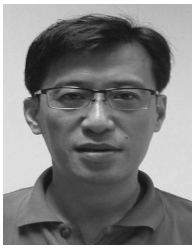
[58] Z. Boulkenafet, J. Komulainen, Z. Akhtar, A. Benlamoudi, D. Samai, S. E. Bekhouche, A. Ouafi, F. Dornaika, A. Taleb-Ahmed, L. Qin, F. Peng, L. B. Zhang, M. Long, S. Bhilare, V. Kanhangad, A. Costa-Pazo, E. Vázquez-Fernández, D. Perez-Cabo, J. J. Moreira-Perez, D. González-Jiménez, A. Mohammadi, S. Bhattacharjee, S. Marcel, S. Volkova, Y. Tang, N. Abe, L. Li, X. Feng, Z. Xia, X. Jiang, S. Liu, R. Shao, P. C. Yuen, W. R. Almeida, F. A. Andaló, R. Padilha, G. Bertocco, W. Dias, J. Wainer, R. da Silva Torres, A. Rocha, M. A. Angeloni, G. Folego, A. Godoy, and A. Hadid, "A competition on generalized software-based face presentation attack detection in mobile scenarios," in *2017 IEEE International Joint Conference on Biometrics, IJCB 2017, Denver, CO, USA, October 1-4, 2017*, 2017, pp. 688–696.

[59] R. Ramachandra and C. Busch, "Presentation attack detection methods for face recognition systems: A comprehensive survey," *ACM Comput. Surv.*, vol. 50, no. 1, pp. 8:1–8:37, Mar. 2017. [Online]. Available: http://doi.acm.org/10.1145/3038924
[60] A. Anjos and S. Marcel, "Counter-measures to photo attacks in face recognition: A public database and a baseline," in *2011 IEEE International Joint Conference on Biometrics, IJCB 2011, Washington, DC, USA, October 11-13, 2011*, 2011, pp. 1–7.
[61] T. F. Pereira, J. Komulainen, A. Anjos, J. M. D. Martino, A. Hadid, M. Pietikäinen, and S. Marcel, "Face liveness detection using dynamic texture," *EURASIP J. Image and Video Processing*, vol. 2014, p. 2, 2014.
[62] 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*, 2013, pp. 1–6.
[63] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," *CoRR*, vol. abs/1803.11097, 2018.
[64] S. Bharadwaj, T. I. Dhamecha, M. Vatsa, and R. Singh, "Computationally efficient face spoofing detection with motion magnification," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2013, Portland, OR, USA, June 23-28, 2013*, 2013, pp. 105–110.
[65] H. Li, W. Li, H. Cao, S. Wang, F. Huang, and A. C. Kot, "Unsupervised domain adaptation for face anti-spoofing," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 7, pp. 1794–1809, 2018.
[66] A. Benlamoudi, K. E. Aiadi, A. Ouafi, D. Samai, and M. Oussalah, "Face antispoofing based on frame difference and multilevel representation," *J. Electronic Imaging*, vol. 26, no. 4, p. 43007, 2017.
[67] K. M. Borgwardt, G. Arthur, M. J. Rasch, K. Hans-Peter, S. Bernhard, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.

Haonan Chen received the B.S. degree from Zhejiang University in 2014. He is currently pursuing the Ph.D.degree with Zhejiang University. His research interest is deep learning, pattern recognition and biometrics (mainly face recognition).

Guosheng Hu is the honorary lecturer of Queens University Belfast and Senior Researcher of AnyVision. He was a postdoctoral researcher in the LEAR team, Inria Grenoble Rhone-Alpes, France from May 2015 to May 2016. He finished his PhD in Centre for Vision, Speech and Signal Processing, University of Surrey, UK in June, 2015. His research interests include deep learning, pattern recognition, biometrics (mainly face recognition), and graphics.

Zhen Lei received the BS degree in automation from the University of Science and Technology of China, in 2005, and the PhD degree from the Institute of Automation, Chinese Academy of Sciences, in 2010, where he is currently an associate professor. He has published more than 100 papers in international journals and conferences. His research interests are in computer vision, pattern recognition, image processing, and face recognition in particular. He served as an area chair of the International Joint Conference on Biometrics in 2014, the IAPR/IEEE International Conference on Biometric in 2015, 2016, 2018, and the IEEE International Conference on Automatic Face and Gesture Recognition in 2015. He is a senior member of the IEEE.



Yaowu Chen received the Ph.D. degree from Zhejiang University, Hangzhou, China, in 1998. He is currently a Professor and the Director of the Institute of Advanced Digital Technologies and Instrumentation, Zhejiang University. His major research fields are embedded system, multimedia system, and networking.

Neil M. Robertson is Professor and Director of Research for Image and Vision Systems in the Centre for Data Sciences and Scalable Computing, at the Queens University of Belfast, UK. He researches underpinning machine learning methods for visual analytics. His principal research focus is face and activity recognition in video. He started his career in the UK Scientific Civil Service with DERA (2000-2002) and QinetiQ (2002-2007). Neil was the 1851 Royal Commission Fellow at Oxford University (2003-2006) in the Robotics Research Group. His autonomous systems, defence and security research is extensive including UK major research programmes and doctoral training centres.

Stan Z.Li received the BEng degree from Hunan University, China, the MEng degree from National University of Defense Technology, China, and the PhD degree from Surrey University, United Kingdom. He is currently a professor and the director of Center for Biometrics and Security Research (CBSR), Institute of Automation, Chinese Academy of Sciences (CASIA). He was with Microsoft Research Asia as a researcher from 2000 to 2004. Prior to that, he was an associate professor in the Nanyang Technological University, Singapore. His research interests include pattern recognition and machine learning, image and vision processing, face recognition, biometrics, and intelligent video surveillance. He has published more than 300 papers in international journals and conferences, and authored and edited eight books. He was an associate editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence and is acting as the editor-in-chief for the Encyclopedia of Biometrics. He served as a program co-chair for the International Conference on Biometrics 2007, 2009, 2013, 2014, 2015, 2016 and 2018, and has been involved in organizing other international conferences and workshops in the fields of his research interest. He was elevated to IEEE fellow for his contributions to the fields of face recognition, pattern recognition and computer vision and he is a member of the IEEE Computer Society.