IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 18, 2019, accepted November 19, 2019, date of publication November 25, 2019,
date of current version December 9, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2955383

A Cascade Face Spoofing Detector Based on Face Anti-Spoofing R-CNN and Improved Retinex LBP

HAONAN CHEN[1], YAOWU CHEN[1,2], XIANG TIAN[3], AND RONGXIN JIANG[1,4]

[1]Institute of Advanced Digital Technology and Instrument, Zhejiang University, Zhejiang 310027, China
[2]Zhejiang Provincial Key Laboratory for Network Multimedia Technologies, Zhejiang University, Zhejiang 310027, China
[3]Zhejiang University Embedded System Engineering Research Center, Ministry of Education of China, Zhejiang 310027, China
[4]State Key Laboratory of Industrial Control Technology, Zhejiang University, Zhejiang 310027, China

Corresponding author: Yaowu Chen (cyw@mail.bme.zju.edu.cn)

This work was supported by the Fundamental Research Funds for the Central Universities.

ABSTRACT In consideration of secure and convenient, face gains increasing attention in variety of fields during the past decades. Since human face is most accessible from our daily life and preserves the richest information, face based biometric systems are widely used in person authentication applications. However, face recognition systems are always challenged by face spoofing attacks. Although, researchers have proposed many face spoofing detection methods, which have achieved great performances, we aim to develop a method to counter face spoofing, which combines the face detection stage and face spoofing detection stage together. In this paper, we design face anti-spoofing region-based convolutional neural network (FARCNN), based on improved Faster region-based convolutional neural network (R-CNN) framework. Motivated by face detection, we regard the face spoofing detection as a three-way classification to distinguish real face, fake face and background. We extend the typical Faster R-CNN scheme by optimizing several important strategies, including roi-pooling feature fusion and adding Crystal Loss function to the original multi-task loss function. In addition, an improved Retinex based LBP is presented to handle the different illumination conditions in face spoofing detection. Finally, these two detectors are further cascaded and achieve promising performances on the benchmark databases: CASIA-FASD, REPLAY-ATTACK and OULU-NPU. Besides, for the purpose of verifying the generalization capacity of the proposed cascade detector, we perform experiments on cross-databases and the results testify the effectiveness of our proposed method.

INDEX TERMS Face spoofing detection, faster R-CNN, crystal loss, Retinex, attention fusion, guided filter.

I. INTRODUCTION

With applications in person authentication with digital devices, biometric-based systems are widely used with different fine-grained biometric cues (e.g., fingerprint, iris, motions and face). With the development of the face recognition during the past decades, face gains increasing attention in different kinds of fields, which is most accessible from our daily life and preserves the richest information. However, in consideration of the information privacy and security, how to counter face spoofing attacks have become an important issue for face-based biometric systems. Specifically, there are four major types of attacks to deal with: printed photo attack, displayed photo attack, replay video attack and 3D mask attack. With the growth of internet and increasing of information leakage, those spoofs are quite easy to obtain, which can be difficult to distinguish from the users' real faces visually. As a result, it is inescapable to develop reliable face anti-spoofing methods and deploy them in the face authorization systems.

Printed, displayed and replay attacks are the most common spoofings and have been well studied from different researchers. In the literature, a variety of different face spoofing detection algorithms have been developed, which have achieved impressive performances on benchmark databases.

To counter the face spoofing attacks, a prejudge is needed for the face recognition systems to find out whether the face is real or not. Thus, the face spoofing detection is treated as a two-category classification problem and the solutions are to learn a classifier which can discriminate between the genuine faces and fake faces effectively. The methods of face anti-spoofing can be divided into three categories in the literature: (1) texture based methods, (2) motion based methods, (3) image quality and reflectance based methods.

The associate editor coordinating the review of this manuscript and approving it for publication was Zahid Akhtar.

For (1), the algorithms develop discriminative texture features to counter various attacks. Most of the methods [1]–[6] design hand-crafted features paying close attention to the texture differences between the real and fake face images, such as LBP and HOG.

In pursuit of more discriminative texture features to adapt to different attack scenarios, Convolutional Neural Network (CNN)-based methods [7]–[11] is employed to learn deep features for face spoofing detection, which exhibit superior performances compared with traditional hand-crafted feature methods. Since training a CNN network needs large quantity of training data, CNN-based methods may confront the risk of overfitting, which may make the performances drop.

For (2), since the still images such as the printed and displayed photos are used for face spoofing attacks, so the basic motions can differentiate a real face from a printed and displayed photo attack. To counter printed and displayed photo attacks, motion based methods (i.e. head motions, eye blinking and lip movements [12]–[15]) are developed. However, the methods may lose efficacy when facing with video replay attacks.

For (3), image quality and reflectance-based methods [16]–[18] are proposed to extract the surface reflection of the different materials and image quality degradation from the real faces and fake faces. Due to the print, displayed and replay attacks can be treated as the photos and videos recaptured from the genuine faces. It is obvious that the recaptured image losed its quality and high frequency information compared genuine ones. Besides, the noise information in the fake face images also can be a significant cue for face anti-spoofing, because of limited resolution and abnormal shading.

In general, the traditional face spoofing detection algorithm includes two stages: detect facial region and then feature extraction and classification with the detected facial region. The traditional face spoofing detection algorithms feed the face region processed by face detection module to the trained detector and output the classification scores of real faces and fake faces. Face detection task is to find the location of faces and output the confidence scores of the regions, which can be regard as a binary classification to differentiate face and background. Motivated by face detection, face spoofing detection task can be regard as a three-way classification for real face, fake face and background. So we aim to explore an architecture to combine the face detection stage and face spoofing detection stage. Region-based CNN (RCNN) [19]–[22] method is one of very important and extremely effective framework in object detection task and face detection task, which can be the countermeasure to solve the two stages of the traditional face spoofing detection at the same time.

Although currently existing face anti-spoofing algorithms have achieved achieve great success, the performance of the detectors are often influenced by external factors such as illumination and image quality. To improve the illumination robustness of the detector, we aim to explore a novel feature to cope with different illumination conditions.

In this work, we designed a face spoofing detection method based on Faster R-CNN framework, called face anti-spoofing R-CNN (FARCNN). This proposed FARCNN combines two stages of traditional face spoofing detection (detect and crop facial region and extract features) into one stage. The original images without cropped are fed into the FARCNN and the bounding boxes and classification scores are outputted directly, which satisfies the realistic application scenarios. In the framework of FARCNN, we improve the existing Faster R-CNN scheme by extending several important strategies, including roi-pooling feature fusion and adding Crystal Loss function to the original multi-task loss function.

In addition, we propose a improved Retinex based LBP to handle the different illumination conditions in face spoofing detection. The traditional Retinex image enhancement algorithm is improved by employing iterative guided filter and luminance components of differernt color spaces for illumination estimation. To make the best use of the enhanced images, we extract LBP features on each component of the enhanced images and concatenate them together. The extracted improved Retinex based LBP features are further fed to Support Vector Machine (SVM) for classification.

In general, the proposed FARCNN is more accurate but light sensitive, while the proposed improved Retinex based LBP is less accurate but light robust. We proposed a standby cascade based on late fusion to take full advantage of these two detectors, which improves the performance of face spoofing detection. Our major contributions are as follows:

- We propose a face anti-spoofing R-CNN (FARCNN) which accepts original face images as the input of the network, which combines the face detection stage and the feature extraction stage together and treat the face spoofing detection as a three-way classification for real face, fake face and background.
- To make the best use of the features of multiple convolution layers, we present an attention-based fusion method to fuse the roi-pooling features adaptively and effectively. Due to the adaptive fusion method, our proposed FARCNN can generate more discriminative features for classification part of the network.
- To minimize the intra-class variations and maximize the inter-class variations of the learned features, we first employ the Crystal loss and center loss for network training for face spoofing detection. The multi-loss function combining Crystal loss and center loss can lead to a better performance compared with traditional softmax loss function.
- To improve the performances of FARCNN in various lighting conditions, a improved Retinex based LBP is proposed. We optimize the existing Retinex algorithm by employing iterative guided filter and luminance components of different color spaces

for illumination estimation. The proposed improved Retinex based LBP features are proved to be discriminative clues for face spoofing detection.

- To take best advantage of the proposed detectors, we present a standby cascade of the FARCNN and the improved Retinex based LBP detector, which improves the illumination robustness of the FARCNN and works better for face spoofing detection.
- To have a fair result comparison with the state of art, we evaluate our proposed cascade detector extensively on benchmark databases of face spoofing detection: CASIA-FASD database, REPLAY-ATTACK database and OULU-NPU database, and have achieved impressive performances. Besides, we also conduct experiments on cross-database and have achieved very competitive results, which verified the great generalization capacity of our proposed cascade detector.

II. RELATED WORKS

A. FACE SPOOFING DETECTION

Since face spoofing detection gains increasing attention recently, researchers have proposed a large quantity of methods in the literature to counter face spoofing. In this section, we briefly review approaches in three categories: texture based methodsïijŇmotion based methods and image quality and reflectance based methods.

1) TEXTURE BASED METHODS

Different texture-based features are explored to counter face spoofing in this category. The features can simply divide into two types: hand-crafted features and CNN-based features.

In [1], the authors discovered the difference of texture between the 2D images and 3D images using the analysis of Fourier spectra and found the different frequency distributions between 2D images and 3D images due to the surface reflection of 2D and 3D surfaces. In [2], the authors extract hidden face texture features using retinex-based and the Difference-of-Gaussian (DoG) filters, which can distinguish real and fake face. In [3], the researchers fed the texture features, which were extracted by multi-scale local binary pattern (LBP), to SVM classifier for real and fake face classification. Reference [4] combined both space and time information and developed a single multiresolution feature extractor to detect face spoofing attacks. Reference [5] proposed the co-occurrence of adjacent local binary pattern (CoALBP) algorithm to extract facial texture features for face spoofing detection. [6] considered the texture in different color spaces and extracted LBP features from HSV and YCbCr color spaces to distinguish fake images from real images. With an increasing variety of face spoofings, one single hand-crafted feature can not meet different needs of face spoofing detection.

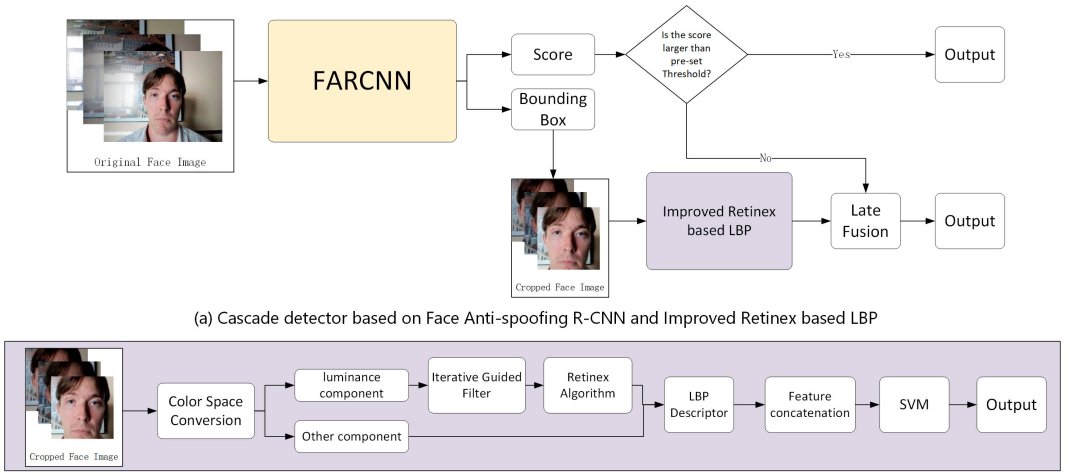
To learn more discriminative features from the data, CNN-based methods are proposed to automatically learn features for face anti-spoofing. In [8], the authors employed pre-trained models on ImageNet and finetuned the model on face-spoofing databases and fed the features into SVM for classification. Reference [7] developed different feature inputs and fed them into CNN network for face spoofing detection. Reference [9] propose a novel two-stream CNN-based approach for face anti-spoofing, which combining hand-crafted feature and CNN-based feature extracted from the face images. These two types of features provide the face image with two liveness scores and the late fusion lead to the final prediction for classification. Reference [23] extracted color local binary patterns (LBP) from the fine-tuned VGG-face model and fed the feature to support vector machine (SVM) for classification. Reference [24] proposed a face PAD method based on the fusion of two complementary attack-specific facial color texture features:RI-LBP and SURF, which also performed well in cross-database evaluation. Reference [25] presented a motion blur analysis-based method and fused 1D convolutional neural network feature and local similar pattern (LSP) feature to detect the replayed video attacks. Reference [10] fused the information of 3D depth shape and rPPG signals to distinguish live and spoof faces. Recently, a depth supervised model was designed in both spatial and temporal domains which is proved to be more robust and discriminative for face spoofing detection in [11].

2) MOTION BASED METHODS

Methods in this category make use of the motions of face such as lips movements, eye blinking and some other motions to counter the photo attacks and display attacks. By modeling different stages of eye blinking, the movement of eyes is used as the effective cue for face anti-spoofing in [12]. In [13], the combination of mouth movement and eye blinking is employed to detect spoofing attacks. Besides, in the IJCB facial spoofing competition [14], some participants used eye blinking and head movement together to distinguish between real faces and spoof faces. For one of the most resent work [15], the researchers enhanced the facial expressions for more robust detection of face anti-spoofing.

3) IMAGE QUALITY AND REFLECTANCE BASED METHODS

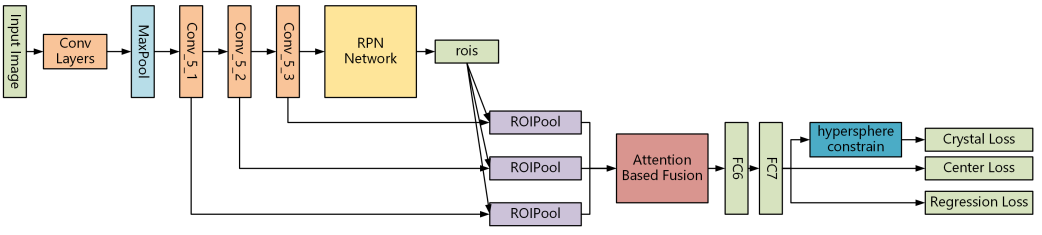
Motivated by the truth that the recapture image and video may cause an image quality drop and a reflectance differences, the methods based on image quality and reflectance are presented in the literature. Reference [16] extracted specular reflection, blurriness, chromatic moment and color diversity on the liquid crystal display (LCD) screen to describe the surface reflection differences between the real and fake faces. Reference [17] fused image-quality features and optical flow features to distinguish live and spoof faces. Moreover, utilizing the noise information of fake face images by the Fourier spectrum, [18] extracted the features which performed well in face spoofing detection. Such image quality and reflectance based methods work well for low-resolution attacks, nevertheless, the performance may drop when facing the high quality spoof artifacts.



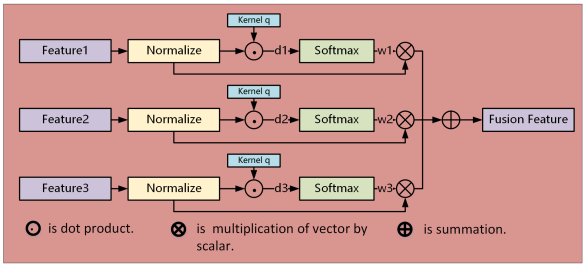
(a) Cascade detector based on Face Anti-spoofing R-CNN and Improved Retinex based LBP

(b) Improved Retinex based LBP Detector

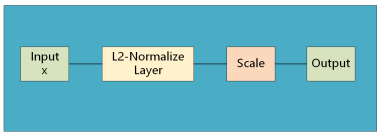
FIGURE 1. **(A) is the overall pipeline; (B) instances the work flow of improved Retinex based LBP detector.**



(A) Attention Based Face Anti-spoofing Region-based Convolutional Neural Network (FARCNN)



(B) Attention Based Fusion Module



(C) L2 Constrain for Crystal Loss

FIGURE 2. **(A) is the pipeline of the Face Anit-spoofing R-CNN; (B) instances the work flow of proposed fusion method based on attention model; (C) clarifies the L2-constraint for Crystal Loss.**

B. REGION-BASED CNNs

With the trend of deep learning, Region-based CNNs (R-CNN) have achieved a dramatical improvement of performances for object detection. The Region-based CNNs methods were initiated by R-CNN [19], which extracts region proposals from the image and then each region of interest (ROI) is classified by a well-trained network. To reduce redundant CNN computations in the R-CNN for speeds-up, the framework has been extended to share the basic convolution features for ROI pooling in [20], [21]. Later, Region Proposal Network (RPN) are presented in Faster R-CNN [22], which achieves further speeds-up compared with Fast R-CNN [21]. Faster R-CNN is extended to many different tasks by the reason of its effectiveness in object detection.

Recently, researchers employ Faster R-CNN to improve the performances of face detection task. Reference [26] proposes a Faster R-CNN based network for face detection, which have a performance rise. Similarly, [27] extends Faster-RCNN to a face detector for multi-task training. CMS-RCNN [28] exploits contextual information to enhance face detect performance. Reference [29] combines two task: face detection and facial keypoint localization together by training a multi-task RPN to gain the performance improvement of these two task.

C. VISUAL ATTENTION MODEL

To focus the perception on the important part of the features, visual attention model can be used to fuse features which offer more information. To fuse color, orientation

and luminance orientation features, [30] proposed a novel fusion method based on attention model, which can simulate the human visual system to make use of the the region of interests of people for performance improvement. In [31], authors proposed an end-to-end CNN network to fuse spatial and temporal features. In addition, attention model has been used in different computer vision tasks, such as image classification [32], emotion recognition [33] and motion recognition [34].

III. METHODOLOGY

Face anti-spoofing is actually regarded as a binary classification problem which is aimed at differentiating real face from fake face. The natural solution of traditional presentation attack detection includes two stages: facial region acquirement and feature extraction. Despite the impressive performances of the traditional CNN-based face spoofing detection algorithms, a model combining face detection and face spoofing detection is needed in practical applications. To put these two stages together, face spoofing detection task can be regard as a three-way classification for real face, fake face and background. We design our face spoofing detection model based on improved Faster R-CNN framework, which has been achieved great success in object detection tasks and face detection tasks.

To adapt to the specific needs of the face spoofing detection, several effective modifications are made in multiple aspects based on Faster R-CNN framework, including roi-pooling feature fusion and adding Crystal Loss function to the original multitask loss function.

Though, FARCNN can achieve promising performances on three benchmark face spoofing databases, FARCNN is sensitive to illumination variations. To handle various lighting conditions, we proposed improved Retinex based LBP and cascade these two detectors to improve the robustness on different lighting conditions.

Generally, we first introduce the framework of our proposed Face anti-spoofing R-CNN. After that, we present our proposed improved Retinex based LBP. Last, we cascade them together with late fusion.

A. FACE ANTI-SPOOFING R-CNN

Similar to Faster RCNN, our Face anti-spoofing R-CNN is a two-stage detector including three major modules: shared convolutional feature layers, region proposal network (RPN) and Fast R-CNN module. The last two modules share the same weight and bias of the convolution feature layers which reduce redundant computations of face spoofing detection task.

The shared basic convolutional layers extract the basic convolutional feature map via multiple convolutional layers and maxpooling layers from an input image. Based on that basic convolutional feature map, the RPN module generates a set of rectangle region proposals which likely contain the faces, called regions of interest (RoIs). The ROIs are processed into fixed-length feature based on the basic convolutional feature map via ROI pooling layer. After that, the ROI pooling features are feed into the Fast R-CNN module for category label prediction (real face vs fake face).

For the RPN module, which is known as an anchor-based method, a set of pre-defined boxes are used for box-classification and box-regression. The typical Faster R-CNN uses anchors which are associate with multiple scales and aspect ratios, aiming to cover various shapes of the bounding boxes. With the prior knowledge that there is only one big face in a face spoofing image, we simplify the scales of the anchors to adapt to this task.

Different with the typical Faster R-CNN, we employ attention model to fuse the multiple convolution layers to make the best use of the ROIs. Besides, we design a multi-task loss function based on the Crystal Loss.

1) FEATURE CONCATENATION

For the typical Faster R-CNN, the features of the region are processed by ROI pooling layer based on the basic convolutional feature map. The RoI pooling features are further fed to the classifier for category label prediction. The RPN module and the RoI pooling layer utilize the same basic features extracted from the convolution layers, which saves a lot of unnecessary calculations. To offer more information of the ROIs, we propose a improved RoI pooling module by combining different layers of the basic feature maps.

In many computer vision tasks, the selection of feature fusion method is importance for improving performance. Inadequate fusion methods may cause the performance of the fusion feature drops compared with individual ones. The commonly used fusion methods contain feature concatenation, feature averaging, feature max pooling and feature min pooling.

In face anti-spoofing task, these traditional methods are proved to be inefficient to maximize the interaction between features of multiple convolution layers. For our solution, we present a fusion method which is based on attention model. Specifically, we first put the features into RoI-pooling and L2-normalize layers before feeding to the proposed attention fusion model.

The proposed attention fusion model is a general structure which is actually regarded as an adaptive-weighted average pooling and can be used in various computer vision tasks. The superiority of attention model can make the fusion features adapt to different task scenarios.

Given a set of features to be fused $\{f_i, i = 1, ..., N\}$, the attention model is trained to generate the fusion weights of the features $\{w_i, i = 1, ..., N\}$ for producing the fusion feature v:

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

Via Eq. (1), learning the the weights $\{w_i\}$ of features is the key of the attention fusion method. When $w_i = 1/N$, the attention fusion becomes traditional feature average

fusion. To adapt to the face spoofing detection, the parameter $N = 3$ and the features are extracted from three convolution layers.

To learn the fusion weights w_i, we employ a kernel q of the same dimensionality of f_i which is only parameters of the model to be trained. he feature vectors are filtered by q via dot product:

$$d_i = q^T f_i \tag{2}$$

$$w_i = \frac{e^{d_i}}{\sum_j e^{d_j}} \tag{3}$$

A vector is generated from the filter, which represent the corresponding feature, named d_i. To further generate the weights w_i subject to $\sum_i w_i = 1$, we feed d_i into the softmax function to generate the weights w_i which is all positive, shown in Eq. (3). From the presentation of attention fusion method, the fusion result r is impertinent with the amount of input features f_i.

2) LOSS FUNCTION

The general face spoofing detection system is trained as a binary classifier which aims to learn to differentiate between the real faces and the fake ones. The binary classifier is commonly trained with softmax loss function, given by Equation (4).

$$L_s = -\frac{1}{M}\sum_{i=1}^{M} log\frac{e^{W_{y_i}^T f(x_i)+b_{y_i}}}{\sum_{j=1}^{C} e^{W_j^T f(x_i)+b_j}} \tag{4}$$

where x_i denotes the i_{th} input feature vector in the batch of size M, $f(x_i)$ indicates the output of the model, y_i is the class label corresponding to the input face image, and W and b are the weights and bias for the layer of classification.

However, when training with the quality imbalance data, the softmax loss is inefficient to model hard negative samples. For our solution, we introduce the Crystal Loss [35], which is newly proposed loss function employed in face recognition tasks. The primary concept of the Crystal Loss is to constrain the feature to a hypersphere with fixed radius by adding an additional L2-constraint to the feature descriptor, given by Equation (5) and (6).

$$\text{minimize} - \frac{1}{M}\sum_{i=1}^{M} log\frac{e^{W_{y_i}^T f(x_i)+b_{y_i}}}{\sum_{j=1}^{C} e^{W_j^T f(x_i)+b_j}} \tag{5}$$

$$\text{subject to } \|f(x_i)\|_2 = \alpha, \quad \forall_i = 1, 2, ...M, \tag{6}$$

where C is the number of the classes and α is the fixed radius on the hypersphere. Similar to softmax loss, the Crystal loss can be coupled with auxiliary loss such as center loss, given by Equation (7).

$$L_c = \frac{1}{2}\sum_{i=1}^{M} \left\|x_i - c_{y_i}\right\|_2^2 \tag{7}$$

For face spoofing detection task, there are only three centers representing real faces, fake faces, and non-faces, respectively. The center loss is very effective in minimizing the intra-class variations, and the Crystal loss as well as the softmax loss have some exploits in maximizing the inter-class variations of the learned features. So, to pursue the discriminative features, we couple Crystal loss with the center loss for the face spoofing detction task. The entire loss function is formulated as:

$$L_{multi} = L_{reg} + \mu L_{center} + \lambda L_{crystal} \tag{8}$$

The regression loss adopted in our method is SmoothL1, which is used for bounding box regression task. The hyper-parameter λ, μ adjust the balancing weights among the three terms of the loss.

Summary In this section, we introduce our proposed face spoofing detection model based on improved Faster R-CNN framework, called FARCNN. The proposed FARCNN is used for two reasons. (1) Different from traditional face spoofing detection, we regard this task as a three-way classification to distinguish real face, fake face and background, which combines the face detection stage and face spoofing detection stage. The performances of face spoofing detection will not be impacted by the effect of face detection module and can take best advantage of the information on full images of the databases. (2) For the application scenarios in reality, the input of the face spoofing detection systems is the original images which is captured by cameras. A face spoofing detection system Combining the face detection task and face spoofing detection task together is satisfied the need of application scenarios in reality.

The feature concatenation with attention model is aim to make the best use of the convolution features from different layers. For traditional Fast RCNN networks, the RoI pooling and the RPN network share the same feature map for generating ROIs and classification, which saving a lot of unnecessary calculations. In order to capture more fine-grained details of the RoIs, we employ attention fusion to obtain the useful information from multiple convolution layers. Specially, we explore which layers should be utilized with some experiments in Section IV-D.

In order to explore a face spoofing detection system for applications, the generalization ability of the method is somehow more important than the performances in intra-database. The center loss is very effective in minimizing the intra-class variations, and the Crystal loss as well as the softmax loss have some exploits in maximizing the inter-class variations of the learned features. Thus, employing Crystal loss and center loss can improve the intra-databse as well as inter-database performances.The specific experiments are shown in Section IV-G.

B. IMPROVED Retinex BASED LBP FEATURE

1) Retinex THEORY

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 propose a

well-designed feature to generalizes better to various environments, mainly various lightings.

Retinex theory was first raised by Land and McCann in 1971 [36]. The Retinex theory is proposed to simulate the human retina system and assumes that the color of the object is determined by its reflection ability of light of different wavelengths, which is independent of the illumination on the object. The source image $S(x, y)$ can be separated into the reflectance image $R(x, y)$ and the illumination image $L(x, y)$, given by Eq.9:

(9):

$$S(x, y) = R(x, y) \cdot L(x, y) \tag{9}$$

where x and y are image pixel coordinates.

According to the theory, illumination image contains the low frequency information, while reflectance image preserves the high frequency information, such as texture and edge. The Retinex Theory is to remove the illumination impact from the source image and gets the reflectance image which can reflect the surface characteristics of the object. To compute the reflectance image, logarithmic transformation deployed on both sides of the Eq.9. So the estimation of $R(x, y)$ can be implemented as follow:

$$log[S(x, y)] = log[R(x, y)] + log[L(x, y)] \tag{10}$$

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.

From Eq.10 we can see, the most vital step of retinex algorithm is to estimate the illumination image. The traditional Retinex based algorithms, such as Single-scale Retinex (SSR) and multi-scale Retinex (MSR), utilize the Gaussian filter for illumination estimation, shown in Eq.11.

$$r(x, y) = s(x, y) - log[S(x, y) * G(x, y)] \tag{11}$$

Symbol '*' is the convolution operation, $G(x, y)$ is Gaussian filter, given as follow:

$$G(x, y) = Ke^{-(x^2+y^2)/c} \tag{12}$$

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

$$\iint F(x, y)dxdy = 1 \tag{13}$$

Summary Traditional Gaussian filter based retinex algorithms have achieve promising performances in image enhancement. However, due to the property of the Gaussian function, there are still some drawbacks for these algorithms.

The retinex theory is based on the assumption that the illumination changes of an image is very slowly, while it may not be tenable in reality. The sharply contrasting area may lead to halo artifacts in enhanced images. In addition, the image filtered by Gaussian filter appears some degree of image blur, which causes the loss of details. To solve these problems, we employ iterative guided filter to estimate the illumination image.

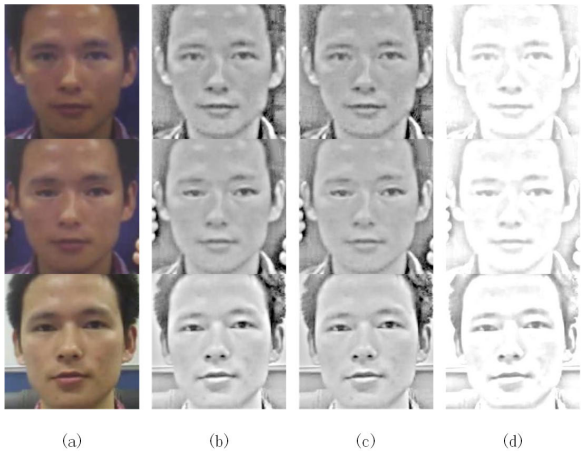


FIGURE 3. Comparison of the SSR algorithm, MSR algorithm and proposed iterative guided filter based retinex algorithm. From top to bottom: print attack, video attack and real face. From left to right: (a) is RGB image, (b) is SSR algorithm result, (c) is MSR algorithm result, (d) is the proposed result.

2) ILLUMINATION IMAGE ESTIMATION

In view of the shortcomings of the traditional Retinex algorithm, iterative guided filter is adopted in this paper instead of gaussian filter to better estimate the illumination image.

Guided filter [37] is an edge-preserving smoothing filter which is fast and non-approximate linear-time algorithm. The guided filter conducts the original image referred to the edges of a guidance image and the guidance image can be the original image or a different image. The filter is a linear model between the guidance image I and the output image q, in addition, the input image p. The definition is as follow:

$$q_i = a_k I_i + b_k, \quad \forall i \in \omega_k \tag{14}$$

where ω_k is the square filter window with the size of $(2r+1)^2$, r is the radius of ω_k, k is the center pixel of the filter window, i is the index of the output and guided image. To compute the linear coefficients a_k and b_k, we minimize the cost function as follow:

$$E(ak, b_k) = \sum_{i \in \omega_k} [(a_k I_i + b_k - p_i)^2 + \varepsilon a_k^2] \tag{15}$$

where ε is a regularization parameter controlling the value of a_k. a_k and b_k are given by linear regression:

$$a_k = \frac{\frac{1}{|\omega|}\sum_{i \in \omega_k} I_i p_i - \mu_k \overline{p}_k}{\sigma_k^2 + \varepsilon} \tag{16}$$

$$b_k = \overline{p}_k - a_k \mu_k \tag{17}$$

where μ_k and σ_k are the mean value and the standard deviation of I in ω_k respectively, $|\omega|$ is the number of the pixels in ω_k, $\overline{p}_k$ is the mean of p in ω_k. The output of the filter can be formulated with a_k and b_k.

$$q_i = \frac{1}{|\omega|}\sum_{i \in \omega_k} (a_k I_i + b_k) = \overline{a}_i I_i + \overline{b}_i \tag{18}$$

where $\overline{a}_i and \overline{b}_i$ is the mean values of a_k and b_k respectively.

FIGURE 4. Scheme one iterative guided filter results on REPLAY-ATTACK databate, given in Eq.20. Average gradient (number in the boxes) showing the smooth degree of each iteration step of guided filter. From left to right, the image are filtered with different ε and fixed r, $r = 4$, $\varepsilon^{(n)} = 0.05^2 \times 2^n$. Samples cover three kind of face image in adverse illumination, from top to bottom: print attack, video attack and real face.

FIGURE 5. Scheme two iterative guided filter results on REPLAY-ATTACK databate, given in Eq.21. Average gradient (number in the boxes) showing the smooth degree of each iteration step of guided filter. From left to right, the image are filtered with different ε and fixed r, $r = 4$, $\varepsilon^{(n)} = 0.05^2 \times 2^n$. Samples cover three kind of face image in adverse illumination, from top to bottom: print attack, video attack and real face.

From the Eq.16 17, when $\sigma_k^2 \gg \varepsilon$, the region is of high variance. So $a_k \approx 1, b_k \approx 0$ and according to Eq.14, $q_i \approx I_i$, which means the filter preserves the edge. On the contrary, when $\sigma_k^2 < \varepsilon$, the pixels in ω_k is similar. So $a_k \approx 0, b_k \approx \mu_k$ and $q_i \approx \mu_k$, which means the filter smooths the region. In particularïijŇthe two parameters: radius r of the square filter window and regularization parameter ε need to be set properly, which may effect the output of the guided filter.

$$G_r = \frac{1}{(M-1)(N-1)} \sum_{h=1}^{M-1} \sum_{j=1}^{N-1} \sqrt{(\Delta I_x^2 + \Delta I_y^2)/2} \qquad (19)$$

where M and N is the size of the image, ΔI_x and ΔI_y is the first-order derivative of the horizontal and vertical directions. The smaller of the gradient, the better smoothing effect.

The selection of these two parameters determines the effect of the illumination estimation. Single scale guided filter can not reach the satisfactory smoothing effect for illumination image estimation, hence, three kinds of schemes for iterative guided filter is proposed in this paper, given by Eq.20, Eq.21 and Eq.22.

$$L^{(n)} = G_F(L^{(n-1)}, L^{(n-1)}) \qquad (20)$$
$$L^{(n)} = G_F(L^{(n-1)}, p) \qquad (21)$$
$$L^{(n)} = G_F(p, L^{(n-1)}) \qquad (22)$$

where $L^{(n)}$ and $L^{(n-1)}$ are illumination images in n_{th} and $(n-1)_{th}$ filter step, p is the input image, $G_F()$ is the guided filter with the parameters: $\varepsilon = \varepsilon^{(n)}$ and $r = r^{(n)}$. Scheme one uses the filtered image as the input image and the guidance image. Scheme two uses the input image as the guidance image to iterate filter, while scheme two employs iterative guided filter to obtain the guidance image and filter the input image with the well-processed guidance image. To decide which scheme is suitable for this task, we conduct experiments employing iterative guided filter with fixed parameter r and multiple sets of ε, shown in Fig.4, 5 and 6. From the average gradient values in the figures, the scheme one outperformed the other schemes, which achieved the best iteration efficiency and effectiveness.

FIGURE 6. Scheme three iterative guided filter results on REPLAY-ATTACK databate, given in Eq.22. Average gradient (number in the boxes) showing the smooth degree of each iteration step of guided filter. From left to right, the image are filtered with different ε and fixed r, $r = 4$, $\varepsilon^{(n)} = 0.05^2 \times 2^n$. Samples cover three kind of face image in adverse illumination, from top to bottom: print attack, video attack and real face.

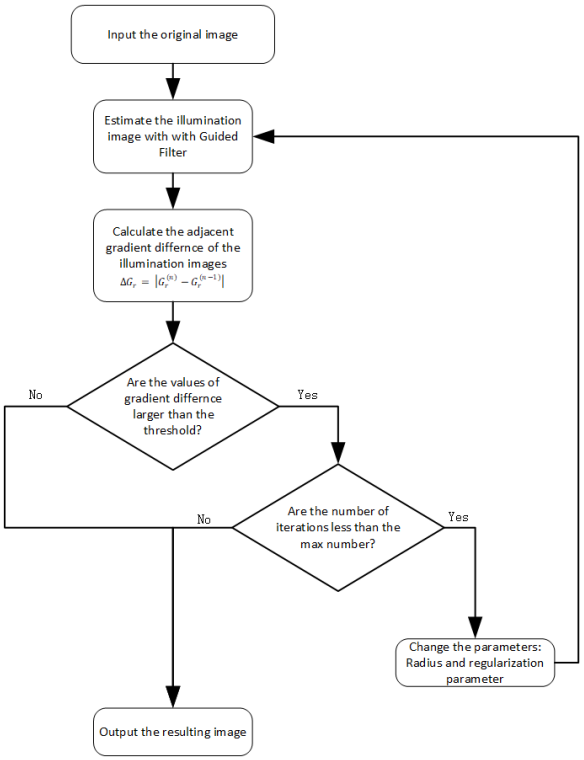


FIGURE 7. The flow chart of the iterative guided filter.

The Eq.20 shows the core mechanism of the iterative guided filter. Guide filtering with scale factor $r^{(n)}$ and regularization parameter $\varepsilon^{(n)}$ is carried out to obtain the illumination image $L^{(n)}$ at step n with illumination image $L^{(n-1)}$ at the previous step, shown in Fig.7. When the adjacent gradient difference of the illumination images (Eq.23) is less than the pre-set threshold ($\Delta G_r < \tau$) or the number of iterations is larger than the pre-set max number ($n < N_{max}$), the iteration process terminates and output the final illumination images.

$$\Delta G_r = \left| G_r^{(n)} - G_r^{n-1} \right| \tag{23}$$

To speed up iteration, we update the parameters in exponent. Specifically, $r^{(n)} = r^{(0)} \times 2^n$ and $\varepsilon^{(n)} = \varepsilon^{(0)} \times 2^n$. For initialization, $r^{(0)} = 2$ and $\varepsilon^{(0)} = 0.05^2$.

The edge preserving and smoothing performance of iterative filtering is reflected in the fact that with the increase of scale parameters and smoothing parameters, various contrast details are gradually smoothed out, which can have a satisfactory estimation of the illumination. The Retinex algorithm results with each iteration step of guided filtered image is shown in Fig.8.

3) IMPROVED Retinex BASED LBP IN DIFFERENT COLOR SPACES

For the traditional Retinex based image enhancement methods, the input RGB image are enhanced separately in three channels, which may lead to severe color distortion. Since RGB color space preserve abundant texture information, the three components of the RGB color space are all closely related to the brightness, that is, as long as the brightness changes, the three components will change accordingly. This motivated us to find appropriate color spaces for illumination estimation.

In this paper, we considered three color spaces: YCbCr, HSL and LAB, to explore the effectiveness for illumination estimation. All of these color spaces have an independent illumination component which can reflect the characteristics of light of a image.

LAB is proposed to describe people's visual perception in a digital way. The 'L' component in the Lab color space is used to represent the brightness of pixels, and its value range is [0,100], representing the range from pure black to pure white. 'A' represents the range from red to green, and the value range is [127,-128]. 'B' represents the range from yellow to blue, and the value range is [127,-128].

YCbCr is commonly used in digital TV and image compression. 'Y' represents the brightness component, 'Cb' is the blue component, and 'Cr' is the red component.

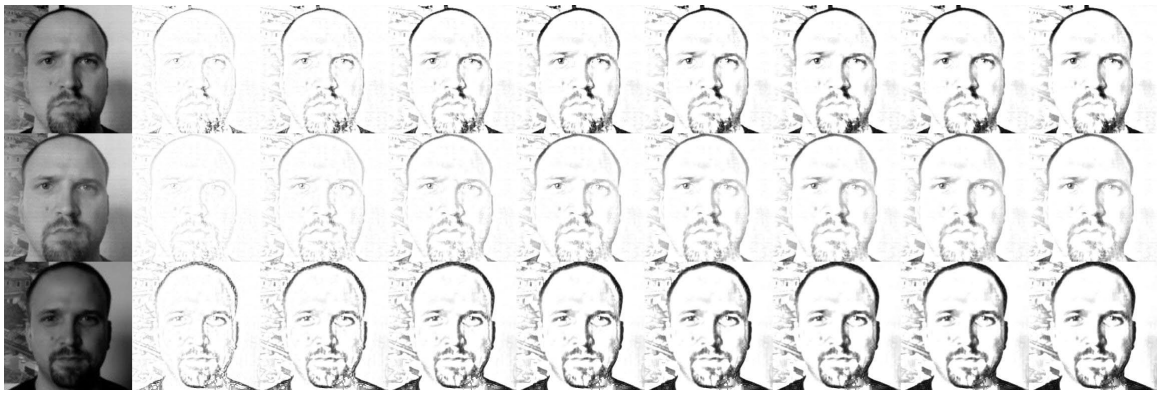


FIGURE 8. Retinex algorithm results based on iterative guided filter on REPLAY-ATTACK database. From top to bottom: print attack, video attack and real face. From left to right: original image, Retinex algorithm results with each iteration step of guided filtered image.

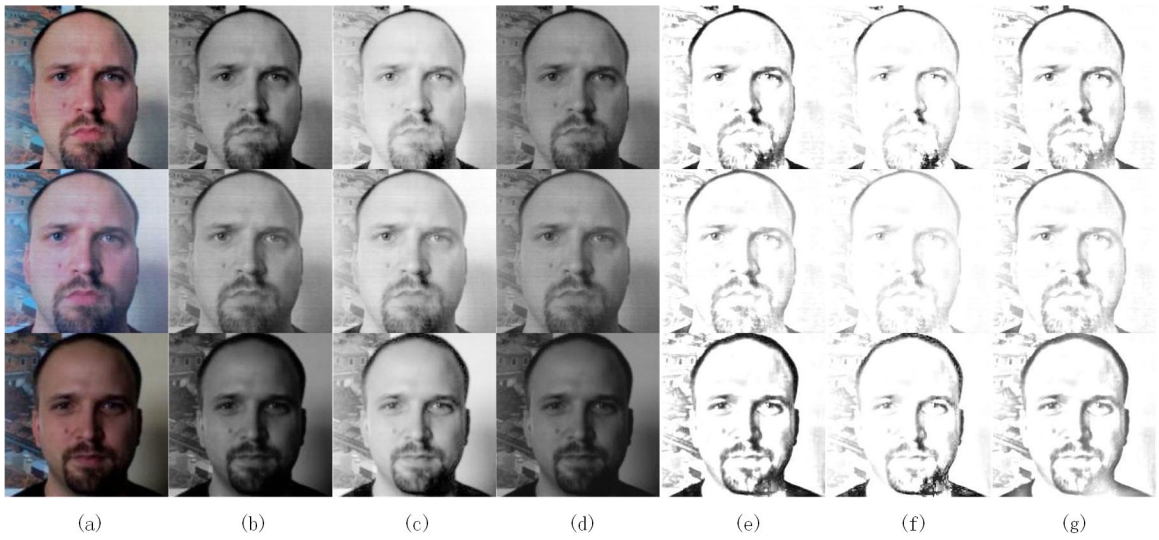


FIGURE 9. Retinex algorithm results based on iterative guided filter with luminance component in different color spaces. From top to bottom: print attack, video attack and real face. From left to right: (a) RGB image, (b) 'L' component in HSL, (c) 'L' component in LAB, (d) 'Y' component in YCbCr, and (e)-(g) is the corresponding iterative guided filter base retinex algorithm results.

HSL is a representation that maps points in the RGB color model into a cylindrical coordinates.'H' is used to represent the fundamental property of color. 'S' refers to the purity of color. The higher the value is, the purer the color is. The lower the value is, the gray gradually becomes. The range ovalue is 0-100%. 'L' represents the brightness component with the range of 0-100%.

To verify the effectiveness of different color spaces for face spoofing detection, firstly, RGB image is converted into LAB, YCbCr and HSL, respectively. And then, 'L' component in LAB, 'Y' component in YCbCr and 'L' component in HSL are employed to estimate illumination images (L_{LAB}, L_{YCbCr} and L_{HSL}). According to the Eq.10, reflectance images can be achieved (R_{LAB}, R_{YCbCr} and R_{HSL}), and the results are presented in Fig.9.

After the acquisition of the reflectance images, we synthetize the reflectance images with the other two color components of these three color spaces. The combinations of the three proposed color spaces are named Retinex-AB, Retinex-CbCr and HS-Retinex. To evaluate the effectiveness of these three proposed color spaces for face spoofing detection, we extract the LBP features in Retinex-AB, Retinex-CbCr and HS-Retinex separately in three channels and concatenate them into three novel LBP: Retinex-AB LBP, Retinex-CbCr LBP and HS-Retinex LBP. The experimental results in three benchmark face spoofing database are presented in Section IV.

C. DETECTOR CASCADE WITH LATE FUSION

The output of the proposed FARCNN is the bounding box of faces and corresponding scores. Despite FARCNN can solve the most cases of face spoofing, when facing the face images in reverse light, the judgment of the classifier becomes uncertain. To face this situation, we develop detector employing improved Retinex based LBP feature as the standby detector.

When the scores outputed by FARCNN are lower than the pre-set threshold (threshold $= 0.9$), these images and the

corresponding bounding boxes will be fed into the proposed improved Retinex based LBP detector. The detector first crops the face images with the bounding boxes and then extracts improved Retinex based LBP features according to the illustration in Section III-B. Finally, we average the scores obtained from improved Retinex based LBP detector and the scores obtained from FARCNN, and make the final decision, shown in Fig.1.

When the scores outputed by FARCNN are higher than the the pre-set threshold, face spoofing judgments from FARCNN is reliable and can be used as the final output. This kind of standby cascade can not only improve the performance of face spoofing detection, but also reduce the calculation.

IV. EXPERIMENTS

In this section, comprehensive experiments have been conducted on three benchmark databases to verify the effectiveness of our method. Firstly, we briefly introduce these three databases of face spoofing detection in Subsection IV-A. Secondly, we provide our experimental settings in Subsection IV-B. Lastly, we present the evaluation results of the two benchmark databases, including inter database results and intra database results.

A. BENCHMARK DATABASE

In this subsection, we give the brief description of CASIA Face Anti-Spoofing Database [38], REPLAY-ATTACK database [39] and OULU-NPU database [40]. These three face spoofing databases contain different kinds of real face images and face spoofing attack images. The brief introductions of CASIA-FASD database, REPLAY-ATTACK database and OULU-NPU database are presented below.

1) THE CASIA FACE ANTI-SPOOFING DATABASE (CASIA FASD)

The CASIA Face Anti-Spoofing Database consists of training set and test set, shown in Fig.10. The face spoofing attacks were generated by recapturing the real face photos and videos with three different cameras. Due to the different capture devices, the attacks contain three imaging qualities: low, normal, and high. In addition, the subjects were required to have some motions in the videos such as eye blink and body sway. The types of the face spoofing attacks are divided into three parts: (1) Warped Photo Attack: The printed face images are warpped to simulates the facial motion of the real people with the resolution of (1920 $\times$ 1080). (2) Cut Photo Attack: The eye regions of a printed face image is cut off and then an attacker exhibits his eyes through the holes of the eye regions to simulate the eye blinking of the real people. Besides, the attacker also exhibits the eyes of a integrated photo through the holes of the eye regions and moving the integrated photo up and down slightly to simulate the eye blinking of the real people. (3) Video Attack: The attacker displays the videos of real people on an iPad and recapture videos via a camera.

FIGURE 10. Sample selected from the CASIA-FASD database. 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.

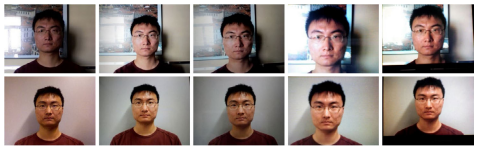


FIGURE 11. Samples selected from the REPLAY-ATTACK database. The first row presents images taken from the controlled scenario, while the second row corresponds to the images from the adverse scenario. From the left to the right: real faces and high definition, mobile and print attacks.

2) REPLAY-ATTACK DATABASE

The REPLAY-ATTACK Database is divided into three subsets: train set, development set and test set, shown in Fig.11. The videos are captured from 50 clients via the webcam on a MacBook with the the resolution of 320 $\times$ 240 and a Canon PowerShot camera and an iPhone 3GS camera with high resolution. The number of video recordings is 1200 in total under two light conditions: controlled condition (captured with a uniform background and light supplied by a fluorescent lamp) and adverse condition (captured with non-uniform background and the day-light). The face spoofing attacks are divided into three parts: (1) Print Attacks: The printed face image are recaptured by different cameras. (2) Mobile Attacks: The real face images and videos are displayed on the screen of an iPhone 3GS and recaptured by cameras. (3) High Definition Attacks: The real face images and videos are displayed on the screen of an iPad and recaptured by cameras.

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.12). These sam were recorded using the front cameras of six mobile devices in three sessions with different illumination conditions and background scenes [40]. The database contains print and video-replay attacks. Two different prints and display devices (Printer 1, Printer 2, Display 1 and Display 2) are employed. The real accesses and attacks are captured by 55 subjects are divided into three subsets

TABLE 1. Comparison between different loss functions on CASIA-FASD database in seven scenarios in terms of EER (%).

Attack Scenarios	Low	Normal	High	Warped	Cut	Video	Overall
Softmax Loss	4.653	3.701	3.260	3.591	3.877	3.154	3.576
Crystal Loss (a=20)	4.572	3.931	2.890	3.209	3.912	3.014	3.566
Crystal Loss (a=25)	4.121	3.851	2.703	4.019	3.831	2.971	3.458
Crystal Loss (a=30)	3.899	3.874	**2.518**	3.654	3.439	2.790	3.309
Crystal Loss (a=35)	3.982	3.750	2.937	3.749	3.219	2.934	3.318
Crystal Loss (a=40)	4.018	3.799	2.879	3.855	3.197	3.006	3.335
Softmax Loss + Center Loss	4.793	**3.611**	3.115	3.675	**3.089**	2.951	3.410
Crystal Loss(a=30) + Center Loss	**3.716**	3.604	2.919	**3.171**	3.190	**2.594**	**3.259**

FIGURE 12. 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.

(20 users) for training, development (15 users) and testing (20 users).

B. EXPERIMENTAL SETTINGS

To compare the results with other algorithms in the state of art, we evaluate our method on benchmark databases following the protocols of these databases. For CASIA-FASD, REPLAY-ATTACK and OULU-NPU, FARCNN is pre-trained with ImageNet and finetuned with the train set of the databases, while improved Retinex based LBP detector is only trained with the train set of the databases.

The results on CASIA-FASD database is reported in terms of Equal Error Rate (EER) and the results on REPLAY-ATTACK database is presented in terms of Equal Error Rate (EER) and Half Total Error Rate (HTER) following commonly used metrics in the literature. Following [41], we evaluate our method on OULU-NPU database with two metrics: Attack Presentation Classification Error Rate (APCER) and Bona Fide Presentation Classification Error Rate (BPCER).

EER is the point where the false rejection rate (FRR) is equal to false acceptance rate (FAR) in receiver operating characteristic (ROC) curve. To achieve HTER, we first find the point of EER and get the threshold corresponding to the EER point on the development set. After that, HTER is given at the point where the threshold computed on the test set is equal to the threshold given by the development set.

TABLE 2. Comparison between different loss functions on REPLAY-ATTACK database in terms of EER (%) and HTER (%).

Methods	EER	HTER
Softmax Loss	0.219	0.752
Crystal Loss (a=20)	0.199	0.570
Crystal Loss (a=30)	0.182	0.437
Crystal Loss (a=40)	0.149	0.312
Crystal Loss (a=50)	0.278	0.838
Crystal Loss (a=60)	0.281	0.794
Softmax Loss + Center Loss	0.195	0.544
Crystal Loss(a=40) + Center Loss	**0.121**	**0.230**

For implementation, we employ VGG-16 as the backbone of our FARCNN, which has been pre-trained on ImageNet. The pre-trained VGG16 model is finetuned on face spoofing detection datasets with the learning rate 0.001 and training epoch 100. Since the sizes of face images in face spoofing databases is quite simple, we use 6 anchors in the RPN module: the size of 256 $\times$ 256, 512 $\times$ 512 and the aspect ratios of 1:1, 1:2, and 2:1. If the IOU of a ROI with any ground truth is greater than 0.5, this ROI is treated as foreground and background otherwise.

For the protocols of CASIA-FASD, the results are presented in seven attacking scenarios. The evaluation of each scenarios of CASIA-FASD is conducted on the data selected from specific protocols separately. As for the REPLAY-ATTACK database, we finetune our FARCNN with the train set and evaluate the model with the development set and test set. Since the database size of the REPLAY-ATACK is larger than CASIA-FASD, we set the different values of α in this experiment. For the evaluation of OULU-NPU database, the results are presented in four protocols.

C. EVALUATION OF DIFFERENT LOSS FUNCTIONS

Loss functions is vital for network training which might lead to the performance improvement of face anti-spoofing. To explore the effect of the different loss functions, we train the network with softmax loss and Crystal loss for various α on CASIA-FASD database and REPLAY-ATTACK database.

For the CASIA-FASD, the softmax loss attains an EER of 3.576%, while Crystal loss achieves the best EER of 3.309% when α is 30, shown in Table 1. From Table 2, we can see Crystal loss ($\alpha = 40$) works better than the results trained with softmax loss in terms of EER

TABLE 3. Comparison between different fusion methods on CASIA-FASD database in seven scenarios in terms of EER (%).

Attack Scenarios	Low	Normal	High	Warped	Cut	Video	Overall
Without Feature Fusion	3.716	3.604	2.919	3.171	3.190	**2.594**	3.259
Feature Concatenate	4.108	3.453	3.155	3.314	**3.167**	2.879	3.241
Feature Average	3.986	**3.344**	3.809	3.212	3.443	2.778	3.402
Feature Max	4.048	3.416	3.017	4.347	4.331	2.969	4.008
Feature Min	3.820	3.438	3.381	3.139	3.747	2.864	3.587
Attention Fusion	**3.516**	3.493	**2.838**	**3.025**	3.206	2.623	**3.049**

(0.149% VS 0.219%) and HTER (0.312% VS 0.752%) on REPLAY-ATTACK database.

To further improve the performance, we couple the Crystal loss as well as softmax loss with center loss and conduct experiments with same experimental conditions. Table 1 lists the results obtained on the CASIA-FASD dataset by different multi-loss functions. From the table, The center loss improves the performance significantly when trained with softmax loss and Crystal loss (Overall:3.410% VS 3.259%), while the results with Crystal loss is better. Besides, with the participation of center loss, the performances are improved on REPLAY-ATTACK database. The network trained with Crystal loss and center loss generally outperform that trained with softmax and center loss in terms of EER (0.121% VS 0.195%) and HTER (0.230% VS 0.544%), shown in Table 2.

From the results above, Crystal loss is efficiently with other auxiliary loss functions and can replace the function of softmax loss in training process. In general, as shown in Table 2 and 1, FARCNN achieve impressive performances both on CAISA-FASD database and REPLAY-ATTACK database.

D. EVALUATION OF DIFFERENT FUSION METHODS

Table 1, Table 2 have verified the effectiveness of the Crystal loss. We further explore the effectiveness of feature fusion strategy in this subsection.

To gain more details of ROIs, we propose to fuse the features of multiple convolution layers to improve the RoI pooling features. Specifically, we fusion the features pooled from $conv3_3$, $conv4_3$, and $conv5_3$ layers. To explore the best fusion method for this task, we conduct the experiments fusing with different fusion methods including attention-based fusion, feature concatenation, feature averaging, feature max pooling and feature min pooling. We compare the results of these fusion methods on different databases separately.

In Table 3, the results of CASIA-FASD are presented in seven scenarios. The attention-based fusion method outperform the other fusion methods with the lowest EER 3.049% in overall scenario, which expresses the effectiveness of our proposed method. Besides, Feature Concatenate and Feature Average achieve the 2nd and 3rd best performances in terms of EER (3.241% and 3.402%).

In Table 4, the results deploying our proposed fusion method achieve better performances compared with the other fusion methods on REPLAY-ATTACK database in terms of EER and HTER (0.093% and 0.026%). The competitive and consistent performances result from the fact that the proposed fusion method can adaptively fuse the different feature layers to adapt to different task scenarios.

TABLE 4. Comparison between different fusion methods on REPLAY-ATTACK in terms of EER (%) and HTER (%).

Methods	EER	HTER
Without Feature Fusion	0.121	0.230
Feature Concatenate	0.119	0.252
Feature Average	0.194	0.407
Feature Max	0.206	0.516
Feature Min	0.180	0.322
Attention Fusion	**0.093**	**0.206**

TABLE 5. Comparison between the attention fusion with different layers on REPLAY-ATTACK in terms of EER (%) and HTER (%).

Methods	EER	HTER
conv5-3	0.142	0.294
conv5-1+conv5-2	0.139	0.254
conv5-2+conv5-3	0.099	0.213
conv5-1+conv5-3	0.105	0.229
conv2-2+conv4-3+conv5-3	0.306	0.461
conv3-3+conv4-3+conv5-3	0.229	0.315
conv4-1+conv4-2+conv4-3	0.160	0.308
conv5-1+conv5-2+conv5-3	**0.093**	**0.206**

To explore the proposed FARCNN should pool from which layers, we conduct more experiments for attention fusion with different layers on REPLAY-ATTACK database, shown in Table 5. Experiments are divided into three parts: (1) pooling from conv-5-3, which is the original scheme of Faster R-CNN; (2) attention fusion with two layers in conv-5; (3) attention fusion with three layers containing lower-level and high-level features. Specifically, features from lower-level convolution layers are scaled to match the scale of high-level features, respectively. From the results of part one and part two, the fusion results are better than the results without fusion. Besides, the fusion involving conv-5-3 layer is more effective than others. As for the results of part three, the performances of multiple level fusion are not satisfactory. Different from object detection and face detection, the sample in face spoofing databases is only contained one single face, which don't need multiple level fusion. In general, the high-level convolution features are vital for this task, which earn largest weights in attention fusion based on the mechanism of attention model. In addition, the amplitudes of features at different layers vary from each other, so it is necessary to normalize the amplitudes with L2-normalization before attention fusion.

TABLE 6. Comparison between different Retinex LBP in three color spaces on CASIA-FASD database in seven scenarios in terms of EER (%).

Attack Scenarios	Low	Normal	High	Warped	Cut	Video	Overall
HS-Retinex LBP	4.257	4.544	4.102	4.051	4.312	4.099	4.154
Retinex-CbCr LBP	4.591	4.329	3.908	4.115	4.708	4.378	4.290
Retinex-AB LBP	4.724	4.651	4.111	4.209	4.800	4.617	4.588
HS-Retinex-CbCr LBP	4.194	4.009	3.790	3.858	4.031	3.608	3.739
HS-Retinex-AB LBP	4.002	4.153	3.632	4.199	4.561	4.033	4.014
Retinex-AB-CbCr LBP	4.018	4.266	3.989	4.301	4.510	4.309	4.281
FARCNN with Attention Fusion	3.516	3.493	**2.838**	3.025	3.206	2.623	3.049
FARCNN + HS-Retinex-CbCr LBP	**2.593**	**3.014**	2.898	**2.431**	**2.581**	**2.011**	**2.359**

E. EVALUATION OF IMPROVED Retinex BASED LBP

The evaluation results above have verified the effectiveness of the FARCNN. We further explore the effectiveness of improved retinex LBP in three color spaces in this subsection.

The extraction of the improved Retinex LBP features is carried out in three steps. Firstly, we first convert the RGB image into HSL, YCbCr and LAB and enhance the luminance component with improved Retinex, presented in Section III-B. Secondly, the enhanced luminance component and the other two color component are processed with LBP descriptor separately. Thirdly, we concatenate the LBP features extracted from these three color component, which are called HS-Retinex LBP, Retinex-CbCr LBP and Retinex-AB LBP for convenience. To further improve the performances of these LBP features, we concatenate the LBP features between different color space and create another three improved Retinex based LBP: HS-Retinex-AB LBP (using luminance component on HSL for Retinex based enhancement), HS-Retinex-CbCr LBP (using luminance component on HSL for Retinex based enhancement) and Retinex-AB-CbCr LBP (using luminance component on YCbCr for Retinex based enhancement). The improved Retinex based LBP features are further fed to SVM for classification.

We conduct an extra experiment to explore the illumination sensitivity of the detectors. the face images in REPLAY-ATTACK database are under two different 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. We evaluate the effectiveness of the detectors on the images on these two illumination conditions separately. From the results in Table 9, FARCNN appears more sensitive for illumination than HS-Retinex-CbCr LBP, in terms of EER and HTER. HS-Retinex-CbCr LBP performs stable, which shows the robustness on strong lightings. After cascade, FARCNN's robustness of illumination is improved and the performances gap between these two illumination conditions shrinks. This indicates that the cascade of FARCNN and HS-Retinex-CbCr LBP can effectively handle various illuminations and achieves better performances.

Table 6 shows the results on CASIA-FASD. From the results, HS-Retinex LBP outperform the other two features (Retinex-CbCr LBP and Retinex-AB LBP), in terms of EER (4.154% vs 4.290% and 4.588%). Besides, HS-Retinex-CbCr LBP attains an EER of 3.739% which is the best among these six improved Retinex based LBP. Though, the performances of the improved Retinex based LBPs are less effective than the performances of FARCNN, the cascade of these two detectors achieves an EER of 2.359%, which improves the performances of FARCNN.

TABLE 7. Comparison between different Retinex LBP in three color spaces on REPLAY-ATTACK in terms of EER (%) and HTER (%).

Methods	EER	HTER
HS-Retinex LBP	2.729	2.993
Retinex-CbCr LBP	2.311	2.359
Retinex-AB LBP	3.871	4.002
HS-Retinex-CbCr LBP	2.145	2.201
HS-Retinex-AB LBP	2.545	2.752
Retinex-AB-CbCr LBP	2.638	2.810
FARCNN + HS-Retinex-CbCr LBP	**0.062**	**0.183**

For the results on REPLAY-ATTACK, shown in Table 7, Retinex-CbCr LBP achieves the best results among single improved Retinex based LBP (HS-Retinex LBP, Retinex-CbCr LBP and Retinex-AB LBP), in terms of EER (2.311%) and HTER (2.359%). HS-Retinex-CbCr LBP outperforms the other two multi-improved Retinex based LBP (Retinex-AB-CbCr LBP and HS-Retinex-AB LBP), in terms of EER (2.145%) and HTER (2.201%). Besides, the cascade of FARCNN and HS-Retinex-CbCr LBP improved the results of FARCNN with EER (0.062%) and HTER (0.183%).

We also conduct evaluations on OULU-NPU database, following [41] and employing four metrics: EER in development set and APCER, BPCER and ACER in test set. Table 8 shows the results of HS-Retinex-CbCr LBP, FARCNN and the cascade of them. For most results in four protocols, the cascade of these two detectors significantly outperforms individual detector.

F. COMPARISONS WITH STATE-OF-THE-ART

To verify the effectiveness of our proposed cascade detector, we compare our results with the state of the art methods for face spoofing detection in Table 10. In general, the proposed method has promising performances compared with other competitors, proving the effectiveness of our proposed cascade detector.

As shown in Table 10, our proposed method achieves the competitive performances in terms of EER (FARCNN+HS-Retinex-YCbCr) on CASIA-FASD database. Besides, our proposed method (FARCNN+HS-Retinex-YCbCr)

TABLE 8. Results achieveing by Retinex LBP, FARCNN and cascade these two detectors on OULU-NPU database in terms of EER (%), APCER (%), BPCER (%) and ACER (%).

Prot.	Methods	Dev	Test		
		EER(%)	APCER(%)	BPCER(%)	ACER(%)
1	HS-Retinex-YCbCr	7.1	10.6	7.2	8.9
	FARCNN with Attention Fusion	1.5	3.8	**6.1**	**4.9**
	FARCNN + HS-Retinex-YCbCr	**1.3**	**3.6**	6.4	5.0
2	HS-Retinex-YCbCr	6.7	7.5	9.7	8.6
	FARCNN with Attention Fusion	2.2	**3.3**	7.1	5.2
	FARCNN + HS-Retinex-YCbCr	**1.6**	3.7	**4.2**	**4.0**
3	HS-Retinex-YCbCr	6.3±1.5	4.5±1.4	9.8±2.1	7.1±1.8
	FARCNN with Attention Fusion	2.1±0.7	4.3±0.4	2.7±1.1	3.5±1.0
	FARCNN + HS-Retinex-YCbCr	**1.8±0.3**	**3.9±2.2**	**2.3±0.9**	**3.1±1.6**
4	HS-Retinex-YCbCr	7.2±0.4	11.5±5.5	10.7±3.6	11.1±4.5
	FARCNN with Attention Fusion	2.1±0.3	13.9±3.1	11.1±3.5	12.5±3.3
	FARCNN + HS-Retinex-YCbCr	**1.8±0.5**	**10.7±3.8**	**10.3±3.8**	**10.5±3.8**

TABLE 9. EER (%) and HTER (%) of FARCNN and HS-Retinex-CbCr LBP detector on adverse illumination and controlled illumination in REPLAY-ATTACK database.

Methods	Adverse illumination		Controlled illumination	
	EER	HTER	EER	HTER
FARCNN	0.103	0.354	0.046	0.153
HS-Retinex-CbCr LBP	2.158	2.469	2.122	2.035
FARCNN + HS-Retinex-CbCr LBP	0.087	0.293	0.039	0.129

TABLE 10. Comparison between our proposed FARCNN and State-of-the-art methods on REPLAY-ATTACK and CASIA-FASD database in terms of EER(%) and HTER(%).

Methods	REPLAY-ATTACK		CASIA-FASD
	EER	HTER	EER
Motion [44]	11.6	11.7	26.6
LBP [39]	13.9	13.8	18.2
LBP-TOP [45]	7.90	7.60	10.00
CDD [46]	-	-	11.8
DOG [47]	-	-	17.0
DMD [48]	5.3	3.8	21.8
IQA [49]	-	15.2	32.4
CNN [50]	6.10	2.10	7.40
IDA [16]	-	7.4	-
Motion + LBP [51]	4.50	5.11	-
vggface + LBP [23]	0.1	0.9	2.3
RILBP + SURF [24]	1.2	4.2	**1.5**
Color-LBP [6]	0.40	2.90	6.20
Bottleneck feature fusion + NN [42]	0.83	**0.00**	5.83
Ours(FARCNN)	0.093	0.206	3.049
Ours(HS-Retinex-YCbCr)	2.145	2.220	3.739
Ours(FARCNN+HS-Retinex-YCbCr)	**0.062**	0.183	2.359

outperforms the other methods in state of the art in terms of EER. In terms of HTER, we achieve the 2nd best performance, slightly lower than [42], while our method performs better than [42] in terms of EER.

For OULU-NPU database, as shown in Table 11, we can achieve a competitive performance for most results under the four protocols. Reference [43] proposed a method using more auxiliary information (3D depth shape and rPPG) and works best.

TABLE 11. 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 [41]	0.6	2.9	10.8	6.9
	GRADANT [41]	1.1	1.3	12.5	6.9
	Depth + rPPG [43]	-	1.6	1.6	1.6
	FARCNN + HS-Retinex-YCbCr	1.3	3.6	6.4	5.0
2	MixedFASNet [41]	1.3	9.7	2.5	6.1
	GRADANT [41]	0.9	3.1	1.9	2.5
	Depth + rPPG [43]	-	2.7	2.7	2.7
	FARCNN + HS-Retinex-YCbCr	1.6	3.7	4.2	4.0
3	MixedFASNet [41]	1.4±0.5	5.3±6.7	7.8±5.5	6.5±4.6
	GRADANT [41]	0.9±0.4	2.6±3.9	5.0±5.3	3.8±2.4
	Depth + rPPG [43]	-	2.7±1.3	3.1±1.7	2.9±1.5
	FARCNN + HS-Retinex-YCbCr	1.8±0.3	3.9±2.2	2.3±0.9	3.1±1.6
4	Massy HNU [41]	1.0±0.4	35.8±35.3	8.3±4.1	22.1±17.6
	GRADANT [41]	1.1±0.3	5.0±4.5	15.0±7.1	10.0±5.0
	Depth + rPPG [43]	-	9.3±5.6	10.4±6.0	9.5±6.0
	FARCNN + HS-Retinex-YCbCr	1.8±0.5	10.7±3.8	10.3±3.8	10.5±3.8

In general, our method has a competitive and stable performance on both CASIA-FASD, REPLAY-ATTACK and OULU-NPU, which shows the superiority of our proposed method.

G. CROSS-DATABASE COMPARISONS

Since CASIA-FASD and REPLAY-ATTACK databases are captured under the different illumination conditions with different cameras, the generalization capacity of face spoofing detection method is important and meritorious. Thus, we evaluate our proposed cascade detector in cross-database protocols between CASIA-FASD database and REPLAY-ATTACK database. To be specific, we train the network on CASIA-FASD database or REPLAY-ATTACK database and evaluate on another database. To quantitatively measure the generalization ability of the proposed method, we employ the metrics HTER which is computed on the development and test sets of the face spoofing databases. Table 12 reports the cross database results and further compares the results with the other methods in the literature.

In Table 12, the performance of our proposed cascade detector drops compared with that train and test on the same database, due to the different imaging conditions of these two databases. Compared with the methods proposed in state of the art, our proposed method (FARCNN + HS-Retinex-YCbCr) performs the 2nd best in terms of HTER (29.4%), when training on REPLAY-ATTACK database and testing on CASIA-FASD database, while [43] is slightly lower than our proposed method (28.4%). When training on CASIA-FASD database and testing on REPLAY-ATTACK database, our proposed method outperformed the other methods in terms of HTER (26.0%).

TABLE 12. Inter-database evaluation results in terms of HTER (%) on the CASIA-FASD and REPLAY-ATTACK database.

Methods	Train	Test	Train	Test
	CASIA FASD	REPLAY ATTACK	REPLAY ATTACK	CASIA FASD
Motion [44]	50.2%		47.9%	
LBP [39]	55.9%		57.6%	
LBP-TOP [45]	49.7%		60.6%	
Motion-Mag [53]	50.1%		47.0%	
Spectral cubes [18]	34.4%		45.5%	
CNN [50]	48.5%		39.6%	
Color-LBP [6]	47.0%		39.6%	
Colour Texture [54]	30.3%		37.7%	
Depth + rPPG [43]	27.6%		28.4%	
Deep-Learning [55]	48.2%		45.4%	
KSA [56]	33.1%		32.1%	
Frame difference [57]	50.25%		43.05%	
Ours(FARCNN)	28.2%		30.9%	
Ours(HS-Retinex-YCbCr)	27.5%		27.1%	
Ours(FARCNN+HS-Retinex-YCbCr)	26.0%		29.4%	

TABLE 13. Inter-database evaluation results for proposed method in terms of maximum mean discrepancy (MMD) on the CASIA-FASD and REPLAY-ATTACK database.

Model	Train	Val	MMD
Baseline + Softmax	CASIA-FASD	CASIA-FASD	0.6891
	REPLAY-ATTACK	REPLAY-ATTACK	0.5973
	CASIA-FASD	REPLAY-ATTACK	1.4597
	REPLAY-ATTACK	CASIA-FASD	1.6971
Baseline + Crystal Loss	CASIA-FASD	REPLAY-ATTACK	1.2782
	REPLAY-ATTACK	CASIA-FASD	1.4902
Baseline + Multi-Loss	CASIA-FASD	REPLAY-ATTACK	1.2982
	REPLAY-ATTACK	CASIA-FASD	1.3988
Crystal Loss + Attention	CASIA-FASD	REPLAY-ATTACK	0.9461
	REPLAY-ATTACK	CASIA-FASD	1.0766
Multi-Loss + Attention	CASIA-FASD	REPLAY-ATTACK	0.8911
	REPLAY-ATTACK	CASIA-FASD	0.9482
HS-Retinex-CbCr	CASIA-FASD	REPLAY-ATTACK	0.7823
	REPLAY-ATTACK	CASIA-FASD	0.8210
FARCNN+HS-Retinex-CbCr	CASIA-FASD	REPLAY-ATTACK	0.7619
	REPLAY-ATTACK	CASIA-FASD	0.8509

In order to find out why the results on cross-database are worse than that on intra database, we introduce maximum mean discrepancy (MMD) [52] to quantitatively indicate the distance between the distributions of training set and testing set.

$$MMD(F_T, F_V) = \left\| \frac{1}{|F_T|} \sum_{f_t \in F_T} \phi(f_t) - \frac{1}{|F_V|} \sum_{f_v \in F_V} \phi(f_v) \right\| \quad (24)$$

As shown in the equation above, $\phi()$ is used to represent the train data features, $f_t \in F_T$ and the validate data features, $f_v \in F_V$. The value of MMD quantifies the domain shift. If the value of MMD is large, the domain shift of the feature distributions is big.

From the result of Table 13, we can see that:

(1) The MMD of intra database is smaller than the value of inter database for proposed methods.

(2) Due to the complex attack types of CASIA-FASD, when we train the model on REPLAY-ATTACK database and test on the CASIA-FASD, the value of MMD is bigger than that we train on CASIA-FASD database and test on REPLAY-ATTACK database.

(3) When training with Crystal loss and Multi-loss, the values of MMD are reduced compared with training with softmax loss.

(4) The fusion of the features extract from different layers with attention model improves generalization capacity compared with the baseline results.

(5) Improved Retinex based LBP detector appears better generalization capacity than FARCNN. The cascade of these two detectors improves generalization capacity of the FARCNN.

V. CONCLUSION

In our wrok, we proposed a cascade face spoofing detector based on face anti-spoofing R-CNN (FARCNN) and improved Retinex based LBP. Our proposed FARCNN employs the effective loss function called Crystal Loss and the fusion method based on attention mechanism to fuse the different feature layer for ROI pooling. Our proposed improved Retinex based LBP uses iterative guided filter for illumination estimation and extracts improved Retinex based LBP feature on different color spaces.

To verify the effectiveness of our method, we evaluated the approach on two challenging databases: CASIA-FASD, REPLAY-ATTACK and OULU-NPU. We achieve the competitive performances in both intra-database and inter-database. Besides, the experiments of loss function show that Crystal Loss can improve the training effect of this task. The proposed attention based fusion method achieves promising performance compared with other traditional fusion methods. The improved Retinex based LBP feature can improve the Illumination robustness.

ACKNOWLEDGMENT

The authors would like to thank the journal reviewers for their valuable suggestions.

REFERENCES

[1] J. Li, Y. Wang, T. Tan, and A. K. Jain, "Live face detection based on the analysis of Fourier spectra," *Proc. SPIE*, vol. 5404, pp. 296–304, Aug. 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 *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 504–517.

[3] J. Määttä, A. Hadid, and M. Pietikäinen, "Face spoofing detection from single images using micro-texture analysis," in *Proc. Int. Joint Conf. Biometrics (IJCB)*, Oct. 2011, pp. 1–7.

[4] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel, "*LBP – TOP* based countermeasure against face spoofing attacks," in *Proc. Asian Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 121–132.

[5] D. Gragnaniello, G. Poggi, C. Sansone, and L. Verdoliva, "An investigation of local descriptors for biometric spoofing detection," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 4, pp. 849–863, Apr. 2015.

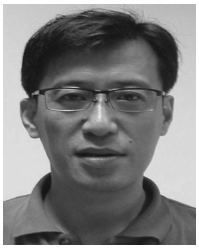
[6] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face anti-spoofing based on color texture analysis," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Quebec City, QC, Canada, Sep. 2015, pp. 2636–2640.

[7] 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 *Proc. 6th Int. Conf. Image Process. Theory, Tools Appl. (IPTA)*, Dec. 2016, pp. 1–6.

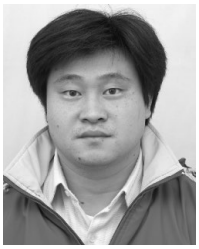
[8] K. Patel, H. Han, and A. K. Jain, "Cross-database face antispoofing with robust feature representation," in *Proc. Chin. Conf. Biometric Recognit.* Cham, Switzerland: Springer, 2016, pp. 611–619.
[9] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu, "Face anti-spoofing using patch and depth-based CNNs," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 319–328.
[10] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 389–398.
[11] Z. Wang, C. Zhao, Y. Qin, Q. Zhou, G. Qi, J. Wan, and Z. Lei, "Exploiting temporal and depth information for multi-frame face anti-spoofing," 2018, *arXiv:1811.05118*. [Online]. Available: https://arxiv.org/abs/1811.05118
[12] G. Pan, L. Sun, Z. Wu, and S. Lao, "Eyeblink-based anti-spoofing in face recognition from a generic webcamera," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.
[13] K. Kollreider, H. Fronthaler, and J. Bigun, "Verifying liveness by multiple experts in face biometrics," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2008, pp. 1–6.
[14] M. M. Chakka *et al.*, "Competition on counter measures to 2-D facial spoofing attacks," in *Proc. Int. Joint Conf. Biometrics (IJCB)*, Oct. 2011, pp. 1–6.
[15] S. Bharadwaj, T. I. Dhamecha, M. Vatsa, and R. Singh, "Face anti-spoofing via motion magnification and multifeature videolet aggregation," Indraprastha Inst. Inf. Technol., New Delhi, India, Tech. Rep. IIITD-TR-2014-002, 2014, pp. 1–14.
[16] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 4, pp. 746–761, Apr. 2015.
[17] L. Feng, L. Po, and Y. Li, "Integration of image quality and motion cues for face anti-spoofing: A neural network approach," *J. Vis. Commun. Image Represent.*, vol. 38, no. 1, pp. 451–460, Jul. 2016.
[18] A. Pinto, H. Pedrini, W. R. Schwartz, and A. Rocha, "Face spoofing detection through visual codebooks of spectral temporal cubes," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4726–4740, Dec. 2015.
[19] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.
[20] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.
[21] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1440–1448.
[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[23] L. Li, X. Feng, X. Jiang, Z. Xia, and A. Hadid, "Face anti-spoofing via deep local binary patterns," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 101–105.
[24] Z. Boulkenafet, J. Komulainen, and A. Hadid, "On the generalization of color texture-based face anti-spoofing," *Image Vis. Comput.*, vol. 77, pp. 1–9, Sep. 2018.
[25] L. Li, Z. Xia, A. Hadid, X. Jiang, H. Zhang, and X. Feng, "Replayed video attack detection based on motion blur analysis," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 9, pp. 2246–2261, Sep. 2019.
[26] H. Jiang and E. Learned-Miller, "Face detection with the faster R-CNN," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May/Jun. 2017, pp. 650–657.
[27] Y. Li, B. Sun, T. Wu, and Y. Wang, "Face detection with end-to-end integration of a ConvNet and a 3D model," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 420–436.
[28] C. Zhu, Y. Zheng, K. Luu, and M. Savvides, "CMS-RCNN: Contextual multi-scale region-based CNN for unconstrained face detection," in *Deep Learning for Biometrics*. Cham, Switzerland: Springer, 2017, pp. 57–79.
[29] D. Chen, G. Hua, F. Wen, and J. Sun, "Supervised transformer network for efficient face detection," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 122–138.
[30] 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, Nov. 1998.
[31] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4747–4756.
[32] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3156–3164.
[33] A. Gupta, D. Agrawal, H. Chauhan, J. Dolz, and M. Pedersoli, "An attention model for group-level emotion recognition," in *Proc. Int. Conf. Multimodal Interaction*, 2018, pp. 611–615.
[34] R. Girdhar and D. Ramanan, "Attentional pooling for action recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 34–45.
[35] R. Ranjan, A. Bansal, H. Xu, S. Sankaranarayanan, J.-C. Chen, C. D. Castillo, and R. Chellappa, "Crystal loss and quality pooling for unconstrained face verification and recognition," 2018, *arXiv:1804.01159*. [Online]. Available: https://arxiv.org/abs/1804.01159
[36] E. H. Land and J. J. McCann, "Lightness and Retinex theory," *J. Opt. Soc. Amer.*, vol. 61, no. 1, pp. 1–11, 1971.
[37] K. He, J. Sun, and X. Tang, "Guided image filtering," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 1–14.
[38] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face antispoofing database with diverse attacks," in *Proc. IAPR Int. Conf. Biometrics*, Mar./Apr. 2012, pp. 26–31.
[39] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Darmstadt, Germany, Sep. 2012, pp. 1–7.
[40] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid, "OULU-NPU: A mobile face presentation attack database with real-world variations," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May/Jun. 2017, pp. 612–618.
[41] Z. Boulkenafet *et al.*, "A competition on generalized software-based face presentation attack detection in mobile scenarios," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 688–696.
[42] 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," *J. Vis. Commun. Image Represent.*, vol. 38, pp. 451–460, Jul. 2016.
[43] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," 2018, *arXiv:1803.11097*. [Online]. Available: https://arxiv.org/abs/1803.11097
[44] A. Anjos and S. Marcel, "Counter-measures to photo attacks in face recognition: A public database and a baseline," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Washington, DC, USA, Oct. 2011, pp. 1–7.
[45] T. F. Pereira, J. Komulainen, A. Anjos, J. M. De Martino, A. Hadid, and M. Pietikäinen, and S. Marcel, "Face liveness detection using dynamic texture," *EURASIP J. Image Video Process.*, vol. 2014, Jan. 2014, Art. no. 2.
[46] J. Yang, Z. Lei, S. Liao, and S. Z. Li, "Face liveness detection with component dependent descriptor," in *Proc. Int. Conf. Biometrics (ICB)*, Madrid, Spain, Jun. 2013, pp. 1–6.
[47] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face antispoofing database with diverse attacks," in *Proc. IAPR Int. Conf. Biometrics (ICB)*, Mar./Apr. 2012, pp. 26–31.
[48] S. Tirunagari, N. Poh, D. Windridge, A. Iorliam, N. Suki, and A. Ho, "Detection of face spoofing using visual dynamics," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 4, pp. 762–777, Apr. 2015.
[49] J. Galbally and S. Marcel, "Face anti-spoofing based on general image quality assessment," in *Proc. 22nd Int. Conf. Pattern Recognit. (ICPR)*, Stockholm, Sweden, Aug. 2014, pp. 1173–1178.
[50] J. Yang, Z. Lei, and S. Z. Li, "Learn convolutional neural network for face anti-spoofing," 2014, *arXiv:1408.5601*. [Online]. Available: https://arxiv.org/abs/1408.5601
[51] J. Komulainen, A. Hadid, M. Pietikäinen, A. Anjos, and S. Marcel, "Complementary countermeasures for detecting scenic face spoofing attacks," in *Proc. Int. Conf. Biometrics (ICB)*, Madrid, Spain, Jun. 2013, pp. 1–7.
[52] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.
[53] S. Bharadwaj, T. I. Dhamecha, M. Vatsa, and R. Singh, "Computationally efficient face spoofing detection with motion magnification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Portland, OR, USA, Jun. 2013, pp. 105–110.
[54] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face spoofing detection using colour texture analysis," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 8, pp. 1818–1830, Aug. 2016.

[55] D. Menotti, G. Chiachia, A. Pinto, W. R. Schwartz, H. Pedrini, A. X. Falc ao, and A. Rocha, "Deep representations for iris, face, and fingerprint spoofing detection," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 4, pp. 864–879, Apr. 2015.
[56] H. Li, W. Li, H. Cao, S. Wang, F. Huang, and A. C. Kot, "Unsupervised domain adaptation for face anti-spoofing," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 7, pp. 1794–1809, Jul. 2018.
[57] A. Benlamoudi, K. E. Aiadi, A. Ouafi, D. Samai, and M. Oussalah, "Face antispoofing based on frame difference and multilevel representation," *J. Electron. Imag.*, vol. 26, no. 4, p. 43007, 2017.

HAONAN CHEN received the B.S. degree from Zhejiang University, in 2014, where he is currently pursuing the Ph.D. degree. His research interests are deep learning, pattern recognition, and biometrics (mainly face recognition).



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 systems, multimedia systems, and networking.



XIANG TIAN received the B.S. and Ph.D. degrees in signal processing from Zhejiang University, Hangzhou, China, in 2001 and 2007, respectively. He is currently an Associate Professor with Zhejiang University. His research focuses on the fields of signal processing and video coding.

RONGXIN JIANG received the B.S. and Ph.D. degrees in computer vision from Zhejiang University, Hangzhou, China, in 2002 and 2008, respectively. He is currently an Associate Professor with Zhejiang University. His major research fields are computer vision and networking.