IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Dual-Path Adaptive Channel Attention Network Based on Feature Constraints For Face anti-spoofing

NANA LI[1], ZHIPENG WENG[1],FANGMEI LIU [1],ZUHE LI [1] AND WEI WANG[2]
[1]College of Computer Science and Technology, Zhengzhou University of Light Industry,Zhengzhou 450002, China
[2]Department of Computing, Xi'an Jiaotong-Liverpool University,Suzhou 215123,China

Corresponding author:WEI WANG(wei.wang03@xjtlu.edu.cn)

This work was supported in part by Henan Provincial Science and Technology Research Project under Grant 242102211020 and 222102210010,the Science and Technology Innovation Project of Zhengzhou University of Light Industry under Grant 23XNKJTD0205, and Jiangsu Science and Technology Programme(General Programme) under Grant BK20221260.

ABSTRACT Interference factors in visible light image data, such as backgrounds and lighting, often lead to poor performance of RGB-based single-modality face anti-spoofing methods. To address these limitations, we propose an innovative face anti-spoofing framework. Within this framework, we design a convolutional neural network (CNN) based on a Dual-path Adaptive Channel Attention (DACA) module, aiming to filter the features of the input facial images to extract key information. In addition, we develop feature constraints method based on Inner Similarity Estimation (ISE), which effectively enhances intra-class consistency by reducing the distance between samples and their class center. This method narrows the intra-class sample distribution and improves class separability, preventing the model from learning excessive irrelevant information and enhancing the robustness and generalization of face anti-spoofing. We test our method on the CASIA SURF dataset, CASIA SURF-CeFA dataset, and CASIA FASD dataset, which shows that our method has significant advantages in distinguishing between live and spoofed faces.

INDEX TERMS face anti-spoofing,attention mechanism,feature constraint,convolutional neural network.

I. INTRODUCTION

WITH the advancement of technology, biometric recognition methods are receiving increasing attention. Due to its convenience, natural interaction, and non-contact nature, face recognition technology [1] has gained wide acceptance and application across various fields, including access control, financial transactions, and security checkpoints. However, the security of face recognition systems is seriously affected by Presentation Attack (PA) methods, which can lead to security risks and even property damage. Face recognition systems are primarily susceptible to three types of PA methods: print attacks [2], video replay attacks [3], and 3D mask attacks [4]. Given the high cost associated with 3D mask attacks, print and video replay attacks remain the most prevalent. Print attacks typically involve printing a face image on paper and using various techniques, such as curling, rotating, or cutting out the eye area, to present the image in front of a live user, attempting to deceive the face recognition system. Video replay attacks involve illegally capturing a user's facial video and playing it through an electronic device to bypass the face recognition system. Although print attacks and video replay attacks are attempting to replicate live facial characteristics, they exhibit distinct differences from genuine faces in terms of texture, movement patterns, and depth features. Based on these differences, a range of anti-spoofing methods can be developed to effectively assess face authenticity. While multi-modality face anti-spoofing can cope with different attacks [5]by fusing various modalities (visible light, infrared, and depth), these methods tend to be more expensive than single-modality methods. In daily life, the devices used for capturing faces are often red, green and blue (RGB) cameras due to their simplicity and cost-effectiveness. Therefore, face anti-spoofing based on RGB data is still of great research significance.

With the significant improvement of computing power, deep learning technology is rapidly emerging in the field of computer vision, demonstrating superior performance in face anti-spoofing compared to traditional methods and greatly enhancing detection accuracy. However, traditional end-to-end deep learning methods for face anti-spoofing still face a lot of challenges in practical applications. Existing deep convolutional neural networks (CNNs) models, when reliant solely on RGB data, may overfit to features that are unrelated to spoofed behaviors, instead of focusing on key cues that distinguish live and spoofed faces. Face anti-spoofing models based on traditional deep learning have demonstrated limited generalization capabilities in the context of complex and varied attack methods.

In order to address these challenges, researchers consider improving the immunity of face anti-spoofing models so that models can adapt to unknown attacks and environmental changes. They are making improvements from two aspects. On the one hand, they improve the model performance by optimizing the CNN model framework. Lucena et al. [6] proposed a face anti-spoofing network, which employs a transfer learning method and optimizes the architecture of the top-level visual geometry group (VGG) network and achieves good results. On the other hand, face anti-spoofing has been considered as a binary classification task. Due to the overlap and confusion between live and spoof face samples in the sample space, it is difficult for the CNNs accurately capture the key features that distinguish the two classes and affect classification accuracy. To overcome this difficulty, researchers have begun to design feature constraints methods based on inner sample estimation. The primary goal of these methods is to narrow the distribution of intra-class samples and minimize confusion and overlap. Hao et al. [7] used contrast loss feature constraints to train twin networks to enhance the differentiation of live faces by driving live sample matching pairs closer while driving non-matching pairs farther away in the sample space. Almeida et al. [8] proposed a multi-target feature constraint to improve the model's sensitivity to different attack methods and device characteristics, while also reducing confusion between different devices, ultimately improving the accuracy of detecting face presentation attacks. Inspired by the above methods, we propose a face anti-spoofing network framework with channel attention and feature constrained learning for RGB-based single-modality images. The main contributions of this paper are as follows:

1) We present a face anti-spoofing framework that combines channel attention and feature constrained learning to improve the performance of single-modality face liveness detection.
2) We construct a dual-path adaptive channel attention (DACA) mechanism by combine two pooling operations with 1D adaptive convolution. The DACA module effectively optimizes the fusion of global and local features, accurately allocates feature weights, and suppresses features unrelated to spoofing cues, which helps the CNN better capture spoofed cues in the face and improves the recognition performance of CNN.
3) We design an inner similarity estimation (ISE) feature constraints based on the distributions of live and spoof samples. The feature constraint prevents the CNN from learning too much interfering information by reducing the distance from the intra-class samples to the class center, which enhances the intra-class consistency while reducing the similarity between samples from different classes.

II. RELATED WORKS

A. TRADITIONAL HANDCRAFTED FEATURE METHODS

In the past, researchers primarily depended on handcrafted features to discriminate the authenticity of faces. Määttä et al. [9] used multi-scale local binary patterns (LBP) to extract the texture features of the image for face live detection. de Freitas Pereira et al. [10] proposed an LBP detection method that focuses on extracting the micro texture and dynamic changes of facial images from three orthogonal planes as a way to distinguish different attacks. Pujol et al. [11] introduced a face anti-spoofing method based on histogram of oriented gradients (HOG) features. However, there are some limitations of handcrafted-based methods such as high cost, low efficiency and weak generalization. In order to improve the detection accuracy, the researchers consider the advantages of combining manual features and deep learning for face authenticity discrimination. Asim et al. [12] combined hand-crafted low-level texture features with high-level spatial and temporal features extracted by CNNs to improve the accuracy of face anti-spoofing. Agarwal et al. [13] used the nonlinear mapping filter of CNN to filter the input image, combined with the histogram features of LBP to compute the convolutional histogram images feature (CHIF) operator, and finally used support vector machine (SVM) to judge the image. Sharifi et al. [14] used overlapping local binary pattern histograms (OVLBP) and VGG16 to extract facial information, then combined the matching scores from both methods to form a fused score vector, which is selectively evaluated to recognize the reality of the face image. Singh et al. [15] proposed a feature fusion model based on LBP and CNN for face anti-spoofing. Although the combination of handcrafted features with deep learning features has achieved some success, it still faces challenges due to the inherent limitations of the former.

B. TRADITIONAL DEEP LEARNING METHODS

The rapid growth of deep learning has led to an increase in the number of detection methods based on this technology. Yang et al. [16] firstly used CNN to extract features and send them into SVM for classification. Shi et al. [17] used ResNet as the backbone and added spatial pyramid pooling (SPP) following the last convolutional layer to break down the extracted feature maps into multiple scales, which effectively utilizes the spatial information present in face images. Ge et al. [18] proposed an innovative model that integrates CNNs with long short-term memory networks (LSTMs). In this model, LSTMs effectively capture long-range dependencies within input sequences, while CNNs focus on extracting localized features from images. However, these methods often ignore the key deception features in the face, which leave significant room for improvement in detection capability. To better capture spoofing features in faces, researchers have adopted attention mechanisms to filter valuable information from feature maps. Alshaikhli et al. [19] used aspatial channel attention module at the spatial and channel levels, respectively, in order to enhance local features while ignoring those unrelated to spoofing cues. Kong et al. [20] integrated the Resnet network with a channel attention mechanism to augment the network's capacity to extract and represent salient features in specific facial regions, such as the nose and cheeks. This approach yielded promising outcomes on RGB data. Sun et al. [21] proposed a network called DatNet, which utilizes a dynamic attention mechanism (Dyattention) to capture spoofing features at different levels. In addition, attention mechanisms can be combined with feature fusion techniques to further improve the accuracy of face anti-spoofing. Chen et al. [22] proposed a two-stream convolutional networks based on attention fusion. The networks fusing RGB and multi-scale retinex (MSR) feature to achieve good results. From the above methods, it is evident that attention mechanisms have significant promise in face anti-spoofing. Moreover, integrating these mechanisms with network models can substantially enhance the accuracy of face anti-spoofing systems.

Due to the interference from lighting variations and backgrounds in RGB face images, some researchers have incorporated feature constrained learning into face anti-spoofing to improve model's stability and capacity for generalization. Chen et al. [23] employed binary focal loss to widen the boundaries between live and spoof samples so that the network can distinguish better. To better identify unknown attacks, Wang et al. [24] suggested a framework named PatchNet, which improves the security of facial recognition systems through fine-grained patch recognition. They proposed feature constraints based on AM-SoftMax loss and self-supervised similarity loss to regulate the patch embedding space. Zheng et al. [25] put forward a joint feature constraint method that combines mean squared error loss (MSE Loss) and symmetry loss to optimize the arrangement of live and spoof face samples in feature space and improves the network's capability to distinguish between live and spoofed faces.

In summary, RGB-based single modality face anti-spoofing suffers from weak generalization ability, low robustness and low anti-interference ability. Therefore, we design the CNN based on dual-path adaptive channel attention to capture more spoofing features. In addition, we propose the feature constraints based on inner similarity estimation, which not only narrows the intra-class distribution by minimizing the distance between intra-class samples and the class center, but also strengths the ability of network to distinguish between categories. Furthermore, by integrating the cross-entropy loss function, we perform joint loss optimization on the network, further improving its generalization performance in RGB-based single-modality face anti-spoofing.

III. METHOD

A. THE OVERALL FRAMEWORK

We propose a new framework named DACN for RGB-based single-modality face anti-spoofing, which combines CNN with channel attention mechanisms to enhance detection capability. we design the ISE feature constraints method, which reinforces intra-class consistency to prevent the CNN from learning excessive features unrelated to spoofing cues.

B. RESNEXT STRUCTURE BASED ON DUAL-PATH ADAPTIVE CHANNEL ATTENTION

Since ResNeXt [26] has achieved good results in the field of face anti-spoofing,we propose a dual-path adaptive channel network (DACN) based on ResNeXt, as shown in Fig.1. In the detection process, the face image is firstly processed by image processing, which randomly divides the face image into multiple image patches and then image patches are sent into the DACN backbone network for feature extraction. The backbone network firstly uses 7 × 7 convolution and downsampling by the maximum pooling layer to capture global features while reducing the size of the feature map. Subsequently, the DACA_Res Block, an enhanced residual block derived from DACA, is employed to extract pivotal features from the downsampled feature maps. The DACA_Res Block contains 32 branches, each reduces the number of input channels from 256 to 4 by a 1 × 1 convolution kernel, which helps to reduce model size and computational burden. Then, a 3 × 3 convolution kernel is used to maintain the channel count while further extracting features. Finally, after restoring the number of channels to 256 using a 1 × 1 convolution kernel, the two parallel paths of the DACA module perform adaptive convolution operations to acquire global and channel features and extract valuable data. After the information extracted from each branch is aggregated, we add the original inputs to the outputs through skip connections, which not only helps the flow of information and reduces the gradient vanishing problem during the training process, but also helps to adaptively adjust the feature weights of each path to enhance important features and suppress irrelevant or redundant information. After the last DACA_Res Block processing, the extracted features are sent to the fully connected layer for classification after pooling operation. Finally, by combining the classification results of each image patch, we can determine whether the image contains fraudulent behavior. Throughout the model's learning and classification process, we adopt a joint optimization based on ISE feature constraints and cross-entropy loss. This joint optimization aims to reduce the gap between similar samples while aggregating samples from different classes and prevent the CNN from learning excessive irrelevant features. In this way, we significantly improve the model's classification accuracy, making it more effective and reliable for practical applications.

C. DUAL-PATH ADAPTIVE CHANNEL ATTENTION

In the last decade, channel attention mechanism has become a key technology a crucial technique to strengthening the performance of CNNs. Since the Squeeze-and-Excitation (SE)

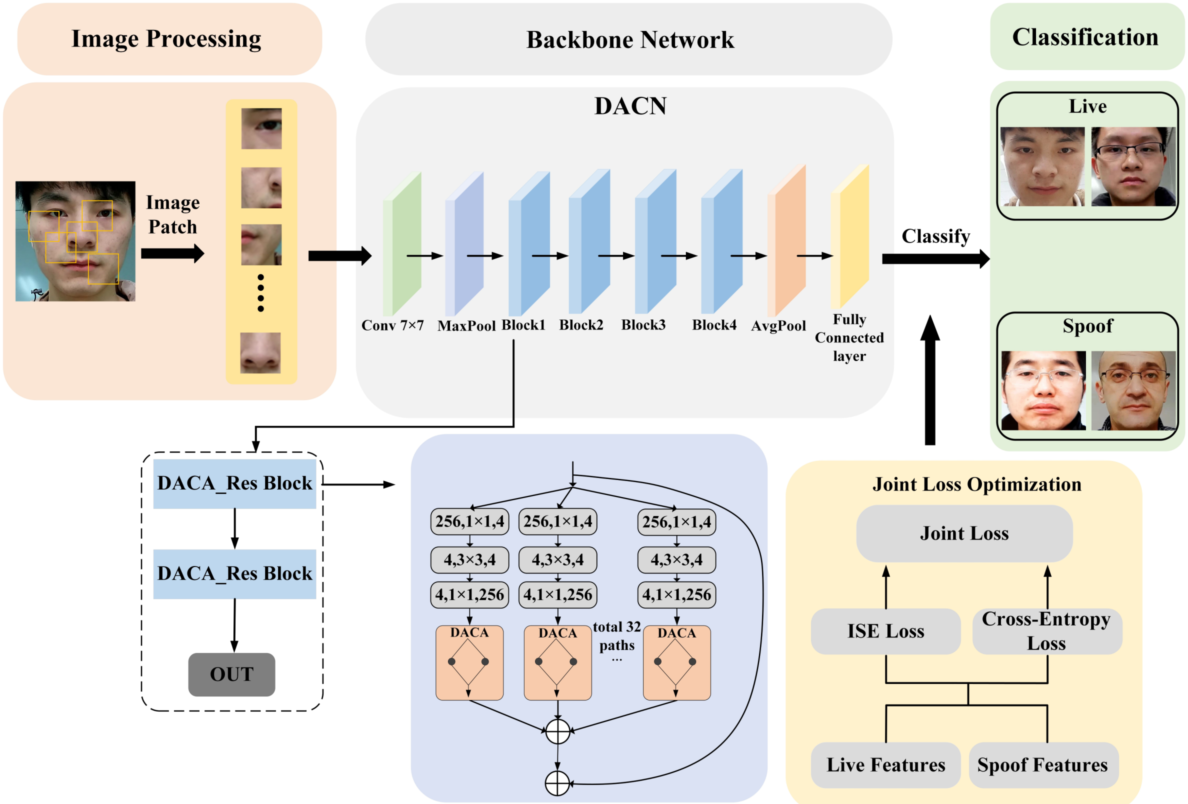


FIGURE 1. **Overall framework of the proposed method. This framework consists of two main components: the dual-path adaptive channel network (DACN) and Joint Loss Optimization. DACN is composed of several DACA_Res blocks, with each DACA_Res Block integrating a residual block and the DACA module. The DACA module effectively captures both global and local channel features from the feature map. Joint Loss Optimization combines cross-entropy loss with ISE feature constraints loss to optimize the network model.**

channel attention mechanism [27] was proposed, it has been widely used in face anti-spoofing, with good results. At the same time, the disadvantage of SE channel attention module is that it uses two fully connected layers to acquire cross-channel interactions. While this reduces complexity of network, but its dimensionality reduction operation negatively affects the prediction of channel attention. This design is inefficient in capturing dependencies between channels and does not effectively integrate global and local channel features, leading to inaccurate feature weight allocation and an inability to effectively learn deceptive cues in facial data. Therefore, to achieve a reasonable allocation of channel feature weights that enables the network to focus more on areas containing significant deceptive cues in faces and to learn these cues efficiently, we design the dual-path adaptive channel attention mechanism (DACA), as shown in Fig.2.

First, we define the input feature map $F(i,j) \in \mathbb{R}^{C\times H\times W}$, where C, H, and W represent the channel size, height, and width, respectively. After the feature map passing through the pooling layer, the spatial features of the feature map are aggregated to generate a channel descriptor, outputting the aggregated feature $U \in \mathbb{R}^{C}$. This process transforms the feature map's size from $C \times H \times W$ to $C \times 1 \times 1$.The feature map $F(i,j)$is sent to two paths for average pooling and max pooling operations, respectively. Max pooling reduces the number of features while preserving those that distinguish live faces from spoof ones. Global average pooling combines the features of each channel, extracting discriminative features for recognizing live and spoofed faces, thus enhancing the network model's generalization capability. After undergoing global average pooling (GAP), we obtain the aggregated feature U_{GAP}. This feature is obtained by averaging the spatial dimensions of the entire feature map across each channel and the process is as follows:

$$U_{GAP} = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} F(i,j) \tag{1}$$

Next, we obtain another set of aggregated features U_{GAP} through the max pooling operation. Max pooling is performed by selecting the maximum value across the spatial dimensions of the entire feature map for each channel, and the process is as follows:

$$U_{GMP} = \max_{1\leqslant i\leqslant H, 1\leqslant j\leqslant M} F(i,j) \tag{2}$$

To efficiently achieve cross-channel interaction, we employ matrix W_K for learning channel attention, defined as follows:

$$\begin{bmatrix} w^{1,1} & \cdots & w^{1,k} & 0 & 0 & \cdots & \cdots & 0 \\ 0 & w^{2,2} & \cdots & w^{2,k+1} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & w^{C,C-k+1} & \cdots & w^{C,C} \end{bmatrix} \tag{3}$$

W_K avoids complete independence between groups after channels have been divided into groups. By employing 1D convolution, we enable all channels to share the same weights, ensuring that W_i not only takes U_i into account but also considers the interactions among its k eighbors. The process is as follows:

$$w_i = \sigma\left(\sum_{j=1}^{k} w^j U_i^j\right) U_i^j \in \Omega_i^k \tag{4}$$

$$w = \sigma(C1D_k(U)) \tag{5}$$

Where σ is the sigmoid function and Ω_i^k epresents the set of the k neighboring channels U_i^j. $C1D_k$ denotes a 1D convolution and its kernel size is k. Another advantage of using 1D convolution is that on the one hand, it avoids the negative impact of dimensionality reduction on learning channel attention. One the other hand, it relies solely on k parameters to capture local cross-channel interactions, ensuring improvements in both efficiency and effectiveness. The size of the convolution kernel k is not a fixed value, but can be adaptively adjusted according to the size of the channel C.The calculation of k is as follows:

$$k = \varphi(C) = \left|\frac{\log_2 C}{\gamma} + \frac{b}{\gamma}\right|_{odd} \tag{6}$$

Where $|t|_{odd}$ indicates the nearest odd number t. In this paper, we set b and γ to 1 and 2, respectively. $\varphi(C)$ is a linear mapping that allows high-dimensional channels to have a longer interaction range, enabling the capture of more global feature information. In contrast, low-dimensional channels have a restricted range, allowing them to focus more on local feature information. The aggregated features U_{GAP} and U_{GMP} are input into 1D convolution for feature filtering and extraction and the output features will be fused. The channel attention wight w_t is obtained through the sigmoid function. The process is as follows:

$$w_t = \sigma\left(C1D_K^1\left(U_{GAP}\right) + C1D_K^2\left(U_{GMP}\right)\right) \tag{7}$$

Where $C1D_K^1$ and $C1D_K^2$ represent the 1D convolution in path 1 and path 2, respectively. Finally, the output feature map $F^{'}$ is obtained by element-wise multiplication with the input feature map F.

$$F' = w_t \otimes F \tag{8}$$

Where F represents the input feature map and $F^{'}$ denotes the output feature map. The DACA processes the feature maps from two paths using 1D adaptive convolution, which not only minimizes the negative effects of dimensionality reduction on channel attention predictions but also reduces computational burden. It flexibly captures channel dependencies across different ranges using 1D adaptive kernels, allowing for a more reasonable allocation of channel feature weights. This enables the network to pay attention to areas of the face that contain abundant deception cues.

D. FEATURE CONSTRAINED LEARNING

Face images contain rich features, including shape, texture, and color. However, when using CNN for anti-spoofing, a common issue is that the network model may overfit the face information, leading to errors in classification tasks. In the space of facial samples, due to the diversity and continuously expanding nature of the samples, the distribution often appears scattered and prone to confusion. This distribution characteristic results in significant overlap regions between live and spoof face samples, which poses challenges for RGB-based anti-spoofing networks. During the training process, this overlap may cause the network to overlook critical deception cues within faces, severely affecting classification performance. To address this issue, we design a feature constraint method called Inner Similarity Estimation (ISE), which minimizes the distance from intra-class samples to the class center and reduces intra-class dispersion and increases inter-class separation. In this way, ISE helps effectively separate live face samples from spoof ones and reduces sample overlap so that network model can learn more features related to deceptive behavior and learn fewer features related to identity. We define the representation of the live sample space Ω_{live} and the spoof sample space Ω_{spoof} as follows:

$$\Omega_{live} = \{\varphi \mid \varphi = x_1, x_2, x_3, \ldots, x_n\} \tag{9}$$

$$\theta_{spoof} = \{\theta \mid \theta = y_1, y_2, y_3, \ldots, y_m\} \tag{10}$$

We find center points from the live face samples and the spoof face samples, denoted as C_x and C_y as follows:

$$C_x = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{11}$$

$$C_y = \frac{1}{m}\sum_{i=1}^{m} y_i \tag{12}$$

After determining the center points, we calculate the distances of the remaining samples from the center points. We denote L_{live} as the feature constraint for live samples and L_{spoof} as the feature constraint for spoof samples, represented as follows:

$$L_{live} = \frac{1}{2}\sum_{i=1}^{n} \|x_i - C_x\|_2^2 \tag{13}$$

$$L_{spoof} = \frac{1}{2}\sum_{i=1}^{m} \|y_i - C_y\|_2^2 \tag{14}$$

Where x_i and y_i the samples in the live class and the spoof class. n and m denote the number of data samples in the live and spoof classes. We combine the feature constraints of

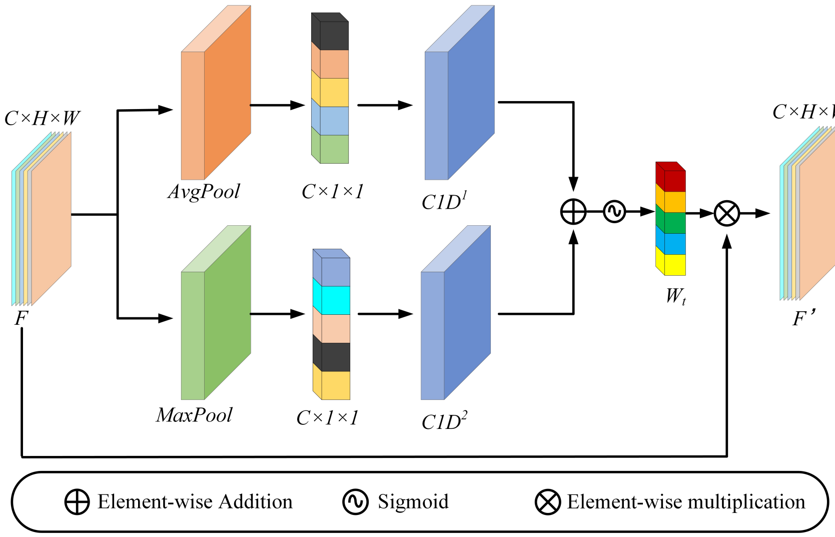


FIGURE 2. The operation process diagram of the Dual-Path Adaptive Channel Attention module.

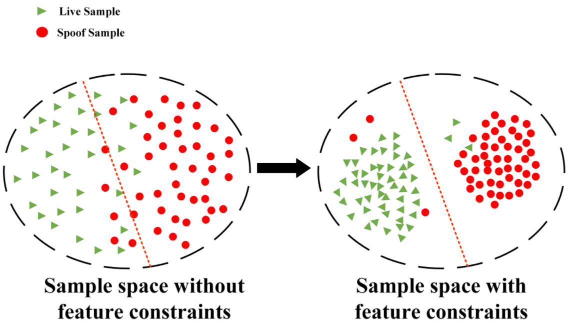

Live Sample
Spoof Sample
Sample space without feature constraints
Sample space with feature constraints

FIGURE 3. Comparison of sample distributions with and without feature constraints from Inner Similarity Estimation (ISE).

live samples and spoof samples to obtain the overall sample feature constraints L_{ISE} as follows:

$$L_{\text{ISE}} = L_{live} + L_{spoof} \tag{15}$$

As shown in Fig.3, without feature constraint learning, the distribution of live and spoof face samples in the sample space is relatively dispersed, with some overlap that makes misclassification likely. After adding the distribution is more compact, and the overlap of them is reduced, which is easy to classify.

E. JOINT LOSS OPTIMIZATION

We optimize the network by using a combined approach of cross-entropy loss and feature constraints to enhance performance during the training and classification process. In each iteration of model training, the cross-entropy loss function gradually narrows the gap between the output and the true labels. This process not only improves the classification accuracy of network but also ensures that the feature representations learned by the network can accurately capture the key information required for the task, while delineating clear boundaries between different categories. As well, ISE feature constraints further strengthen the model's generalization capability by directing it to prioritize learning features essential for distinguishing between categories while minimizing attention to irrelevant details. This method enables the model to classify more stably and reliably when faced with diverse data. The process of joint loss optimization is illustrated in Fig.4. The cross-entropy loss function can be expressed as:

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{16}$$

Where N is the number of samples, y_i is the true label for the ith class and $\hat{y}_i$ is the predicted likelihood for the ith class by the network model. The joint loss is defined as follows:

$$L_{\text{Joint}} = \alpha L_{CE} + \beta L_{ISE} \tag{17}$$

Where L_{CE} is the cross-entropy loss function and L_{ISE} is the ISE feature constraints loss. Our proposed framework operates as shown in Algorithm 1.

IV. EXPERIMENTS

A. DATASETS

We conducted experimental tests on three benchmark datasets: CASIA-SURF dataset [28],CASIA-SURF CeFA cross-ethnicity dataset [29] and CASIA-FASD dataset [30].

CASIA-SURF:CASIA-SURF is a large face anti-spoofing dataset containing multiple modalities designed to support research in face recognition and fraud prevention techniques. As shown in Fig.5, The dataset covers three modalities of data, RGB images, depth images and infrared images, collected from 1000 individuals with a total of 21,000 video samples. Each sample contains one real video clip and six different attack video clips to simulate diverse fraud scenarios in real applications. The dataset is divided into training, validation, and testing sets, containing 300, 100, and 600

Algorithm 1 The train process of DACN model

```
DATA: The mixed live and spoof dataset $\Omega = \{x_i, y_i\}_{i=1}$
Initialize: CNN model $\varphi_0(\cdot)$, $L_{ISE}$ – ISE feature constraints loss, $L_{CE}$ – cross-entropy loss.
for epoch = 1 to epoch_nums do
    Shuffle live data samples $\{x_i \mid i = 1, 2, \ldots, n\}$ and spoof data samples $\{y_i \mid i = 1, 2, \ldots, m\}$
    Compute the center of $x_i$ and compute feature constraint loss $L_{\text{live}}$ based on (13).
    Compute the center of $y_i$ and compute feature constraint loss $L_{\text{Spoof}}$ based on (14).
    Compute the overall sample feature constraints $L_{ISE}$ based on (15).
    Compute the binary loss $L_{CE}$ of predicted values and label values based on (16).
    Compute the joint loss $L_{\text{Joint}} = \alpha L_{CE} + \beta L_{ISE}$.
    Update model parameters.
end for
Evaluate $\varphi_F(\cdot)$ on the testing data.
OUTPUT: Trained model parameters $\varphi_F(\cdot)$.
```

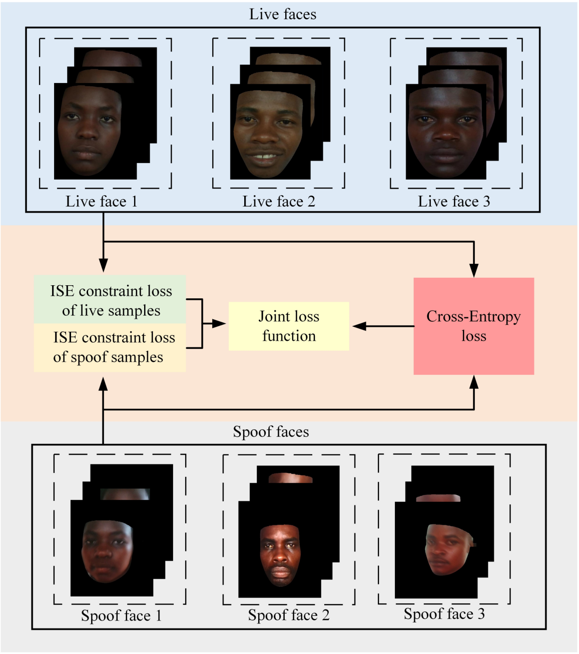


FIGURE 4. the process of joint loss optimization.

subjects, respectively, and providing 148,000, 48,000, and 295,000 frames of video data to ensure the adequacy and diversity of model training and evaluation.

CASIA-SURF CeFA:CASIA-SURF CeFA dataset consists of a 2D attack subset and a 3D attack subset. As shown in Fig.6, the 2D attack subset includes 2D attack samples collected from 1,500 subjects in the Americas, East Asia, and Central Asia. The data types include RGB visible light images, depth images, and infrared images, with a total of 4,500 live samples and 13,500 attack samples, amounting to 18,000 samples in total. The 3D attack subset contains 5,538 3D attack samples collected from 107 subjects, including 5,364 mask attack samples collected from 99 subjects under six different lighting conditions, and 192 samples of mustache or glasses attacks collected from eight subjects under four different lighting conditions. All samples for 3D attacks are stored in video format. The 2D attack subset is designed with five protocols and a total of 12 sub-protocols. Each ethnicity's 500 subjects are divided into three non-overlapping subsets. Each protocol includes three data subsets: the training set, validation set, and test set, containing 200, 100, and 200 subjects, respectively.

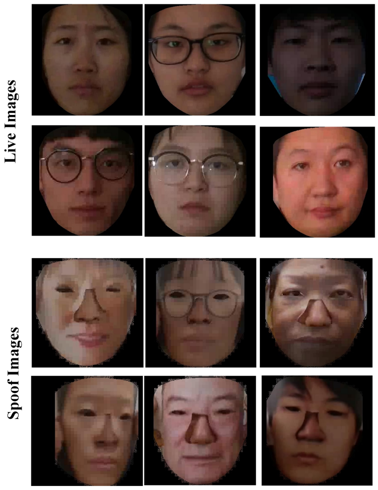


FIGURE 5. Examples of the CASIA-SURF dataset.

CASIA-FASD:CASIA-FASD dataset records the live access and spoofed attack behaviors of 50 different subjects, comprising approximately 5,123 live images and 7,534 spoofed images. The spoofed faces are created from high-quality recordings of live faces. As shown in Fig.7, the facial images are classified into three quality categories: low, medium, and high. Furthermore, the dataset incorporates three distinct categories of synthetic facial images, namely warped photo attack (Bending printed images), cut photo attack (Cutting out certain parts of the photo), and video attack.

B. EVALUATION METRIC

There are three evaluation metrics are employed: attack presentation classification error rate (APCER), normal presentation classification error rate (NPCER), and average classification error rate (ACER). In the case of the CASIA-FASD dataset, equal error rate (EER) is used as the evaluation metric. APCER and NPCER indicate classification error rates, while ACER represents the average of APCER and NPCER.

$$TPR = \frac{TP}{TP + FN} \tag{18}$$

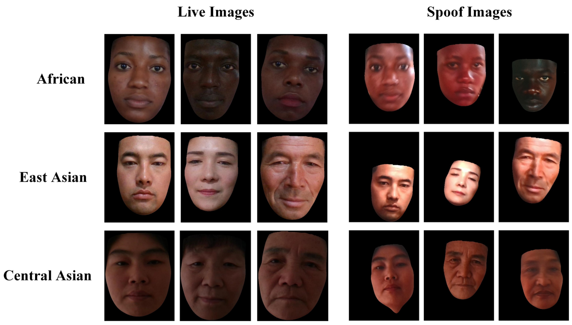


FIGURE 6. Examples of the CASIA-SURF CeFA dataset.

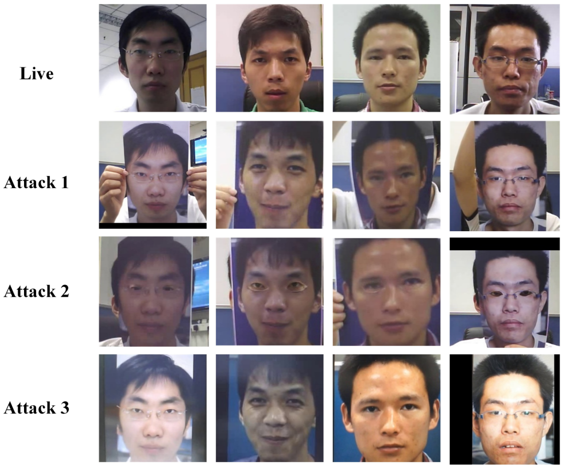


FIGURE 7. Examples from the CASIA-FASD dataset.

$$FPR = \frac{FP}{FP + TN} \tag{19}$$

$$APCER = \frac{FP}{FP + TN} \tag{20}$$

$$NPCER = \frac{FN}{FN + TP} \tag{21}$$

$$ACER = \frac{APCER + NPCER}{2} \tag{22}$$

$$EER = \frac{TPR + FPR}{2} \tag{23}$$

Where TP, FP, TN and FN represent the true positive, false positive, true negative and false negative.

C. TRAINING SETTING

We constructed DACN by using the PyTorch deep learning framework and trained it on an NVIDIA 4090. The size of the face image is 112 × 112, and the whole training process is divided into 10 theories with 50 iterations per round. We used a stochastic gradient descent algorithm with the initial learning rate set to 0.1 and momentum and weight are set to 0.9 and 0.0005, respectively. α and β in (17) are both set to 1.

V. RESULTS AND ANALYSIS

A. EXPERIMENTAL EVALUATION

1) CASIA-SURF Dataset

We input image patches of three different sizes (48×48, 32×32, and full image) into the network for training and tested on the CASIA-SURF dataset. We compared our performance with other methods, including Spatial and Channel Attention [31], SPP [17], Large-scale Multimodal [28], ResNext-50 [26], and TTN-s [32]. Table 1 shows that our proposed method achieves the lowest APCER of 1.67%, the lowest NPCER of 1.57%, and the lowest ACER of 2.42%. These evaluations demonstrate that our method effectively distinguishes between live and spoofed faces when dealing with single-modality RGB images.

TABLE 1. Classification results of each protocol on the CASIA-SURF CeFA dataset.

Protocol	Method	APCER (%)	NPCER (%)	ACER (%)
Spatial and channel attention[31]	Full Image	5.2	2.6	3.9
SPP [17]	Full Image	–	–	6.4
SE-Net [29]	Full Image	1.74	4.21	2.97
Large-scale multimodal[28]	Full Image	8.0	14.5	11.3
ResNext-50 [26]	Full Image	21.76	15.06	18.41
TTN-S [32]	16×16	3.8	3.2	3.5
Ours	32×32	5.89	**1.57**	3.73
Ours	48×48	**1.67**	3.17	**2.42**
Ours	Full Image	2.77	3.13	2.95

2) CASIA SURF-CefA Dataset

The 2D attack subset of the CASIA SURF-CeFA dataset contains five protocols and a total of 12 sub-protocols. We tested on the challenging Protocol 4 and its three sub-protocols (Protocol 4_1, Protocol 4_2, and Protocol 4_3). At the same time, we compare our method with PSMM-Net [29], SD-Net [33], and CDCN [34].Table 2 presents the experimental results of our proposed framework for each protocol. The ACER score on Protocol 4_1 is 2.13%, the ACER score on Protocol 4_2 is 2.73%, and the ACER score on Protocol 4_3 is 1.08%. We achieve the lowest NPCER of 0.56% in Protocol 4_1 and the lowest APCER of 0.45% in Protocol 4_3. These results prove that our proposed network structure performs exceptionally well in the RGB data, effectively addressing face spoofing attacks across different ethnicities and attack methods, demonstrating strong generalization ability.

TABLE 2. Classification results of each protocol on the CASIA-SURF CeFA dataset.

Protocol	Method	APCER (%)	NPCER (%)	ACER (%)
Protocol 4_1	PSMM-Net [29]	5.0	3.3	4.2
	SD-Net [33]	5.72	18.5	12.11
	CDCN [34]	11.17	2.5	6.83
	Ours	3.7	**0.56**	2.13
Protocol 4_2	PSMM-Net [29]	7.7	9.0	8.4
	SD-Net [33]	7.33	11.25	9.29
	CDCN [34]	6.67	2.0	4.33
	Ours	0.76	4.70	2.73
Protocol 4_3	PSMM-Net [29]	10.8	4.3	7.6
	SD-Net [33]	3.17	27.0	15.08
	CDCN [34]	3.72	3.0	4.33
	Ours	**0.45**	1.71	**1.08**

3) CASIA-FASD Dataset

Table 3 presents the results of our proposed method compared to other methods on the CASIA FASD dataset. Our proposed method achieves the lowest *EER* is 2.10%, significantly outperforming other methods. This shows that our proposed network has good robustness against false face attacks with different qualities.

TABLE 3. Classification results of RGB images on the CASIA-FASD dataset

Method	EER(%)
CNN [16]	4.92
DPCNN [35]	4.50
Patch CNN [36]	2.37
LSTM-CNN [37]	5.17
DeepPixel [38]	2.60
LiveNet [39]	4.59
Ours	**2.10**

B. ABLATION ANALYSIS

To demonstrate the effectiveness of our designed DACN and feature constrained learning, we conducted ablation experiments on the CASIA SURF dataset.

1) Impact of channel attention mechanism

To investigate the impact of the attention mechanism on the classification performance of the network model, we test the performance of ResNeXt-50 with and without DACA module under three input patch sizes: 32×32, 48×48, and full image. As shown in Table 4, the ACER of the ResNeXt-50 without the DACA module is 4.90%, 4.72%, and 12.19%, respectively. The ACER of the ResNeXt-50 with the DACA module decreased to 2.62%, 2.58%, and 3.78%. In addition, the ResNeXt-50 with the DACA module achieves the lowest APCER is 3.16% and the lowest NPCER is 0.20%. These results indicate that the DACA module helps the network model learn deceptive cues in faces more effectively, significantly improving the model's accuracy in live detection.

In addition, we also compare the effects of adding the CBAM module and the SE module to the ResNeXt-50 network. Table 5 shows that our proposed attention mechanism effectively extracts key features from face images, significantly reducing the classification error rate.

TABLE 4. Classification performance with and without the DACA module.

Protocol	Method	APCER (%)	NPCER (%)	ACER (%)
ResNeXt-50 [40]	32×32	7.19	2.61	4.90
	48×48	6.97	2.47	4.72
	Full Image	12.46	11.92	12.19
ResNeXt50+DACA(Ours)	32×32	**3.16**	2.08	2.62
	48×48	4.96	**0.20**	**2.58**
	Full Image	3.18	4.38	3.78

TABLE 5. Classification performance with different attention mechanisms.

Method	APCER (%)	NPCER (%)	ACER (%)
ResNeXt-50+CBAM [41]	4.26	4.00	4.15
ResNeXt-50+SE [27]	2.34	5.24	3.80
ResNeXt-50+DACA(Ours)	4.96	**0.20**	**2.58**

2) Impact of feature constrained learning

We optimized the network using a combination of Inner Similarity Estimation (ISE) feature constraints and cross-entropy loss. To investigate the effectiveness of the feature constraints, we test the ResNeXt-50 with cross-entropy loss and with both ISE feature constraints and cross-entropy loss. Table 6 shows that the ACER for the ResNeXt-50 model using only cross-entropy loss is 4.90%, 4.72%, and 12.19% for input patch sizes of 32×32, 48×48, and Full Image, respectively. In contrast, the ACER for the ResNeXt-50 with both ISE feature constraints and cross-entropy loss is 4.39%, 3.85%, and 4.67%, respectively. With the Patch Size of 32×32, we get the lowest NPCER is 0.98% after adding ISE feature constraints. The experimental results demonstrate that the combined optimization of the network with feature constraints and cross-entropy loss helps the model reduce the learning of facial features unrelated to spoofing cues, decrease the classification error rate and improve the accuracy of face anti-spoofing.

TABLE 6. The classification performance with and without Inner Similarity Estimation (ISE).

Method	Patch Size	APCER (%)	NPCER (%)	ACER (%)
	32 × 32	7.19	2.61	4.90
ResNeXt-50	48 × 48	6.97	2.47	4.72
	Full Image	12.46	11.92	12.19
	32 × 32	7.80	**0.98**	4.39
ResNeXt-50+ISE(Ours)	48 × 48	4.38	3.32	**3.85**
	Full Image	**3.27**	6.07	4.67

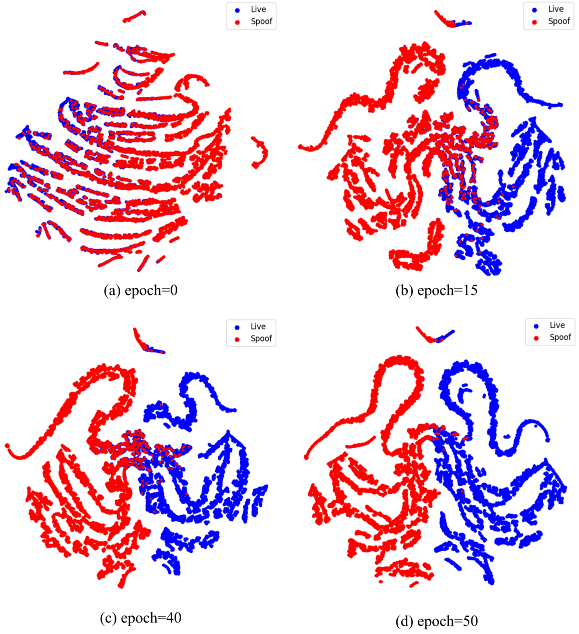
C. VISUALIZATION AND ANALYSIS

1) Feature constraint visualization

By applying the t-SNE algorithm for visual analysis of the sample space, we observe significant changes in the sample distribution at different training stages, as illustrated in Fig.8. At epochs 0, 15, 40, and 50. At epoch 0, live and spoof samples are almost indistinguishable in the feature space, with a high degree of overlap, making them difficult to separate. At epoch 15, we begin to notice a trend where the two classes of samples are aggregating towards opposite sides. Although some overlap remains, this change indicates that the model is learning to differentiate between the two classes. At epoch 40, the overlap between the samples is further reduced, and most samples can be clearly distinguished, each forming relatively concentrated clusters. This demonstrates that the model's discriminative ability is gradually improving. At epoch 50, the distribution of live and spoof samples becomes very distinct, with a clear boundary between the two, forming two separate and well-defined regions. This result strongly proves that as training progresses, the model's discriminative ability is significantly enhanced. The input samples, which were initially mixed, gradually aggregate towards opposite sides and eventually separate, forming two relatively independent regions. This shows that our proposed joint optimization method effectively improves the its discriminative capability.

2) Attention Visualization

To gain deeper insights into the performance of our proposed dual-efficient channel attention mechanism in face anti-spoofing, we employ Grad-CAM activation functions for visualization to generate attention heatmaps. Fig.9 shows a comparison between the attention heatmaps of the original ResNeXt-50 model and our improved model when recognizing live and spoofed faces. The red regions represent areas where the model is highly focused, possibly containing important face spoofing cues. The yellow regions indicate areas to which the model pays attention, though they are slightly less important than the red regions; and the blue regions represent areas where the model considers there to be little to no spoofing cues, with lower importance.



(a) epoch=0
(b) epoch=15
(c) epoch=40
(d) epoch=50

FIGURE 8. Sample distribution at different epochs.

From the attention heatmap, it can be seen that ResNeXt-50 model pays considerable attention to the image background, while its focus on the face region is relatively low. In contrast, our proposed network architecture is able to focus more precisely on the key parts of the face. This indicates that, first, the DACA module effectively assists the CNN model focus more accurately on the facial region, allowing for more effective feature extraction. Second, our designed network can identify the key areas of the face where potential spoofing cues may exist (such as the eyes, mouth, and nose), while reducing attention to areas unrelated to spoofing.Our proposed network is able to flexibly adjust its focus when faced with different types of attacks. For instance, in the case of an attack using a handheld printed photo,For instance, in the case of an attack using a handheld printed photo, the model pays more attention to a human hand within the photograph.

VI. CONCLUSION

In this paper, we propose an RGB-based single-modality face anti-spoofing framework on DACN and ISE feature constraints. The DACA module helps the CNN focus more on areas of the face containing numerous spoofing cues, rather than on areas like the image background that contain fewer cues, allowing for better extraction of spoofing information. During training, we improve the network's performance by applying ISE feature constraints in conjunction with cross-entropy loss. The feature constraints aim to reduce the similarity among samples within the same class while increasing the distinction between samples from different classes. This makes the sample distribution in the sample space more compact and easier to classify, while minimizing the CNN's learning of distracting information in faces. Experimental

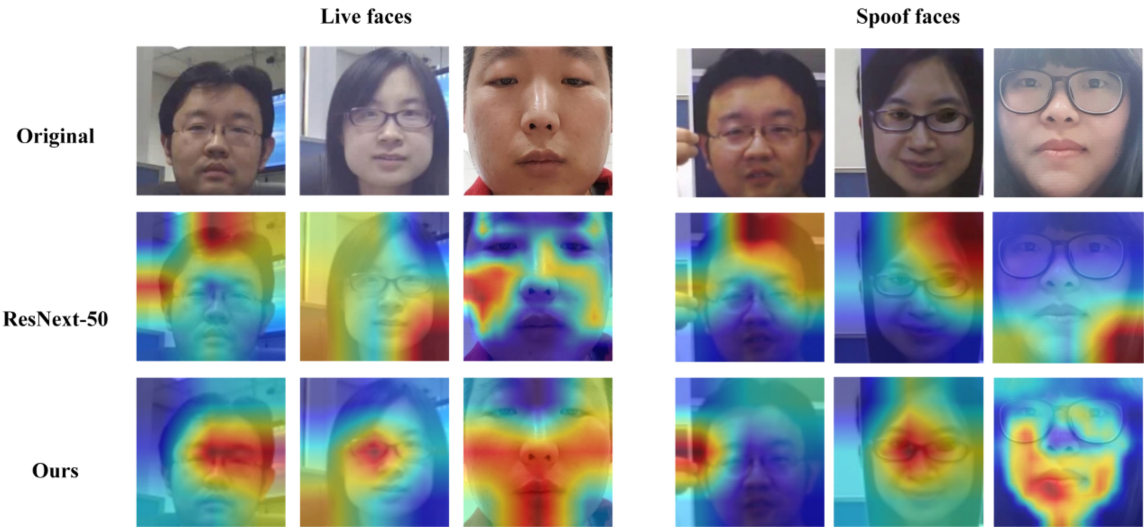


FIGURE 9. Attention visualization of our proposed method.

results indicate that the proposed method effectively extracts spoofing cues from faces, leading to significant performance enhancements on established face anti-spoofing benchmarks.

REFERENCES

[1] Z. Yu, A. Liu, C. Zhao, K. H. Cheng, X. Cheng, and G. Zhao, "Flexible-modal face anti-spoofing: A benchmark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.Workshops. (CVPRW)*, 2023, pp. 6346–6351.

[2] X. Yu, X. Huang, X. Ye, B. Liu, and G. Hua, "Multimodal proxy-free face anti-spoofing exploiting local patch features," *IEEE Signal Process. Lett.*, vol. 31, pp. 1695–1699, 2024. Doi: 10.1109/LSP.2024.3418710.

[3] Y. Liu and X. Liu, "Spoof trace disentanglement for generic face anti-spoofing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3813–3830, 2023, doi: 10.1109/TPAMI.2022.3176387.

[4] L. Birla, P. Gupta, and S. Kumar, "Sunrise: Improving 3d mask face anti-spoofing for short videos using pre-emptive split and merge," *IEEE Trans. Dependable Secure Comput.*, vol. 20, no. 3, pp. 1927–1940, 2023, doi: 10.1109/TDSC.2022.3168345.

[5] S.-Q. Liu, X. Lan, and P. C. Yuen, "Multi-channel remote photoplethysmography correspondence feature for 3d mask face presentation attack detection," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 2683–2696, 2021, doi: 10.1109/TIFS.2021.3050060.

[6] O. Lucena, A. Junior, V. Moia, R. Souza, E. Valle, and R. Lotufo, "Transfer Learning Using Convolutional Neural Networks for Face Anti-spoofing," in *Image Analysis and Recognition*, 2017, vol. 10317, pp. 27–34.

[7] H. Hao, M. Pei, and M. Zhao, "Face Liveness Detection Based on Client Identity Using Siamese Network," in *Pattern. Recognit. Comput. Vis.*, 2019, vol. 11857, pp. 172–180.

[8] W. R. Almeida, F. A. Andaló, R. Padilha, G. Bertocco, W. Dias, R. d. S. Torres, J. Wainer, and A. Rocha, "Detecting face presentation attacks in mobile devices with a patch-based cnn and a sensor-aware loss function," *PLOS ONE*, vol. 15, no. 9, pp. 1–24, Sep. 2020, doi: 10.1371/journal.pone.0238058.

[9] J. Määttä, A. Hadid, and M. Pietikäinen, "Face spoofing detection from single images using micro-texture analysis," in *Int. Joint. Conf. Biom. (IJCB)*, 2011, pp. 1–7, doi: 10.1109/IJCB.2011.6117510.

[10] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel, "Lbp- top based countermeasure against face spoofing attacks," in *Computer Vision-ACCV 2012 Workshops: ACCV 2012 International Workshops, Daejeon, Korea, November 5-6*. Springer, 2013, pp. 121–132.

[11] F. A. Pujol, M. J. Pujol, C. Rizo-Maestre, and M. Pujol, "Entropy-based face recognition and spoof detection for security applications," *Sustainability*, vol. 12, no. 1, p. 85, 2019.

[12] M. Asim, Z. Ming, and M. Y. Javed, "Cnn based spatio-temporal feature extraction for face anti-spoofing," in *Int. Conf. Image, Vis. Comput. (ICIVC)*, 2017, pp. 234–238, doi: 10.1109/ICIVC.2017.7984552.

[13] A. Agarwal, M. Vatsa, and R. Singh, "Chif: Convoluted histogram image features for detecting silicone mask based face presentation attack," in *Proc. Int. Conf. Biometrics Theory, Applications Systems. (BTAS)*, 2019, pp. 1–5, doi: 10.1109/BTAS46853.2019.9186000.

[14] O. Sharifi, "Score-Level-based Face Anti-Spoofing System Using Handcrafted and Deep Learned Characteristics," *IJIGSP*, vol. 11, no. 2, pp. 15–20, Feb. 2019, doi: 10.5815/ijigsp.2019.02.02.

[15] R. P. Singh, R. Dash, and R. K. Mohapatra, "LBP and CNN feature fusion for face anti-spoofing," *Pattern Anal Applic*, vol. 26, no. 2, pp. 773–782, 2023, doi: 10.1007/s10044-023-01132-4.

[16] J. Yang, Z. Lei, and S. Z. Li, "Learn Convolutional Neural Network for Face Anti-Spoofing," Aug. 2014, arXiv:1408.5601.

[17] L. Shi, Z. Zhou, and Z. Guo, "Face anti-spoofing using spatial pyramid pooling," in *Int. Conf. Pattern. Recognit. (ICPR)*, 2021, pp. 2126–2133, doi: 10.1109/ICPR48806.2021.9412407.

[18] H. Ge, X. Tu, W. Ai, Y. Luo, Z. Ma, and M. Xie, "Face anti-spoofing by the enhancement of temporal motion," in *Int. Conf. Advances. Comput. Technol. Inf. Sci. Commun. (CTISC)*, 2020, pp. 106–111, doi: 10.1109/CTISC49998.2020.00025.

[19] M. Alshaikhli, O. Elharrouss, S. Al-Maadeed, and A. Bouridane, "Face-fake-net: The deep learning method for image face anti-spoofing detection : Paper id 45," in *Eur. Workshop. Visual. Inform. Process. (EUVIP)*, 2021, pp. 1–6, doi: 10.1109/EUVIP50544.2021.9484023.

[20] Y. Kong, X. Li, G. Hao, and C. Liu, "Face Anti-Spoofing Method Based on Residual Network with Channel Attention Mechanism," *Electronics*, vol. 11, no. 19, p. 3056, Sep. 2022, doi: 10.3390/electronics11193056.

[21] C.-Y. Sun, S.-L. Chen, X.-J. Li, F. Chen, and X.-C. Yin, "Danet: Dynamic attention to spoof patterns for face anti-spoofing," in *Inte. Conf. Pattern Recognit. (ICPR)*, 2022, pp. 1929–1936, doi: 10.1109/ICPR56361.2022.9956725.

[22] H. Chen, G. Hu, Z. Lei, Y. Chen, N. M. Robertson, and S. Z. Li, "Attention-based two-stream convolutional networks for face spoofing detection," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 578–593, 2020, doi: 10.1109/TIFS.2019.2922241.

[23] B. Chen, W. Yang, H. Li, S. Wang, and S. Kwong, "Camera invariant feature learning for generalized face anti-spoofing," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 2477–2492, 2021, doi: 10.1109/TIFS.2021.3055018.

[24] C.-Y. Wang, Y.-D. Lu, S.-T. Yang, and S.-H. Lai, "Patchnet: A simple face anti-spoofing framework via fine-grained patch recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 20 249–20 258, doi: 10.1109/CVPR52688.2022.01964.

[25] W. Zheng, M. Yue, S. Zhao, and S. Liu, "Attention-based spatial-

temporal multi-scale network for face anti-spoofing," *IEEE Trans. Biom. Behav. Identity Sci.*, vol. 3, no. 3, pp. 296–307, 2021, doi: 10.1109/TBIOM.2021.3066983.

[26] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5987–5995, doi: 10.1109/CVPR.2017.634.

[27] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141, doi: 10.1109/CVPR.2018.00745.

[28] S. Zhang, X. Wang, A. Liu, C. Zhao, J. Wan, S. Escalera, H. Shi, Z. Wang, and S. Z. Li, "A dataset and benchmark for large-scale multi-modal face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 919–928, doi: 10.1109/CVPR.2019.00101.

[29] A. Liu, Z. Tan, J. Wan, S. Escalera, G. Guo, and S. Z. Li, "Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing," in *Proc. IEEE. Win.Conf. Appl Comput Vis. (WACV)*, 2021, pp. 1178–1186, doi: 10.1109/WACV48630.2021.00122.

[30] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face antispoofing database with diverse attacks," in *Inte. Conf. Biom.(ICB)*, 2012, pp. 26–31, doi: 10.1109/ICB.2012.6199754.

[31] G. Wang, C. Lan, H. Han, S. Shan, and X. Chen, "Multi-modal face presentation attack detection via spatial and channel attentions," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. (CVPRW)*, 2019, pp. 1584–1590, doi: 10.1109/CVPRW.2019.00200.

[32] Z. Wang, Q. Wang, W. Deng, and G. Guo, "Learning multi-granularity temporal characteristics for face anti-spoofing," *IEEE Trans. Inf. Forensics Secur.*, vol. 17, pp. 1254–1269, 2022, doi: 10.1109/TIFS.2022.3158062.

[33] A. Liu, X. Li, J. Wan, Y. Liang, S. Escalera, H. J. Escalante, M. Madadi, Y. Jin, Z. Wu, X. Yu, Z. Tan, Q. Yuan, R. Yang, B. Zhou, G. Guo, and S. Z. Li, "Cross-ethnicity face anti-spoofing recognition challenge: A review," *IET Biometrics.*, vol. 10, no. 1, pp. 24–43, Jan. 2021, doi: 10.1049/bme2.12002.

[34] Z. Yu, Y. Qin, X. Li, Z. Wang, C. Zhao, Z. Lei, and G. Zhao, "Multi-modal face anti-spoofing based on central difference networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.Workshops. (CVPRW)*, 2020, pp. 2766–2774, doi: 10.1109/CVPRW50498.2020.00333.

[35] Y. Sun, H. Xiong, and S. M. Yiu, "Understanding deep face anti-spoofing: from the perspective of data," *The Visual Computer.*, vol. 37, no. 5, pp. 1015–1028, 2021.

[36] 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 *International Conference on Image Processing Theory, Tools and Applications. (IPTA)*, 2016, pp. 1–6, doi: 10.1109/IPTA.2016.7821013.

[37] Z. Xu, S. Li, and W. Deng, "Learning temporal features using lstm-cnn architecture for face anti-spoofing," in *Asian Conf.Pattern Recognit. (ACPR)*, 2015, pp. 141–145, doi: 10.1109/ACPR.2015.7486482.

[38] A. George and S. Marcel, "Deep pixel-wise binary supervision for face presentation attack detection," in *Inte. Conf. Biometrics (ICB)*, 2019, pp. 1–8, doi: 10.1109/ICB45273.2019.8987370.

[39] Y. A. U. Rehman, L. M. Po, and M. Liu, "LiveNet: Improving features generalization for face liveness detection using convolution neural networks," *Expert Systems with Applications*, vol. 108, pp. 159–169, Oct. 2018, doi: 10.1016/j.eswa.2018.05.004.

[40] A. George and S. Marcel, "Learning one class representations for face presentation attack detection using multi-channel convolutional neural networks," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 361–375, 2021, doi: 10.1109/TIFS.2020.3013214.

[41] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis*, 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2$_1$.

NANA LI received the master's degree in communication and information systems from Beijing University of Posts and Telecommunications,in 2006.She has been engaged in teaching and research work with the School of Computer Science and Technology, Zhengzhou Institute of Light Industry, since May 2006, where she is currently an Associate Professor and the Master's Tutor. Her research interests include machine learning,image processing and deep learning.

ZHIPENG WENG received the B.S. degree in network engineering from the Zhengzhou University of Light Industry in 2023, where he is currently pursuing the master's degree in computer technology with the College of Computer Science and Technology. His research interests include image processing and deep learning.

FANGMEI LIU received her master degree in computer applied technology from Zhengzhou University of Light Industry in 2007. She is a lecturer at Zhengzhou University of Light Industry. Her research interests include information integration, data processing and deep learning.

ZUHE LI received the Ph.D. in Information and Communication Engineering from Northwestern Polytechnical University in 2017. He is an associate professor at the Zhengzhou University of Light Industry. His research interests include computer vision and deep learning.

WEI WANG received his Ph.D. in Computer Science from the University of Nottingham in 2009. He is a senior associate professor at the Department of Computing, Xi'an Jiaotong-Liverpool University, China. His research interests lie in the broad area of data and knowledge engineering and deep learning.