scientific reports

Check for updates

OPEN

A high-performance adaptive fusion network for face anti-spoofing detection

Hui Qi[1,3,4,7], Rui Han[1,7], Kaige Duan[5,7], Ying Shi[1,2], Xiaobo Qi[1], Chenghai Gao[4]✉ & Lifang Ren[6]

In view of the limitations of current cross-domain face liveness detection models in generalization ability and deep feature representation, this paper proposes a high-performance adaptive fusion network for face anti-spoofing detection. This method innovatively introduces the face depth map fusion mechanism, combines the ResNet-18 backbone network to extract common features in multiple source domains, strengthens the distinguishable feature capture between real faces and spoofing attacks. Meanwhile, a content feature extraction architecture of "dynamic convolution + bottleneck attention module" is designed. It is combined with adaptive instance normalization and central difference convolution for collaborative optimization, breaking through the bottleneck of insufficient representation of depth details in traditional feature extraction. Finally, through adversarial training of domain discriminators, dual alignment of multi-source domain features and categories is achieved, effectively alleviating the problem of cross-domain data distribution differences. A large number of training and test results conducted on the four benchmark datasets of OULU-NPU, MSU-MFSD, CASIA-FASD and ReReplay Attack show that the performance of the proposed method is significantly better than that of the existing algorithms, providing a more innovative technical path for cross-domain face liviality detection.

Face liveness detection[1] is a biometric security technology that distinguishes real faces from fake ones (such as photos, videos, 3D printed masks, etc.) through technical means. Its core objective is to prevent identity fraud attacks and ensure the security of face recognition systems. It usually uses facial physiological features (such as blinking, facial micro-expressions) or multimodal information (such as infrared imaging, depth data) to determine "whether it is a real living person", for example, verifying the specific behavioral responses of a living person through dynamic instructions (such as "opening the mouth", "turning the head"), or capturing the temperature distribution differences of real skin through an infrared camera. This technology is widely applied in scenarios such as financial payment, access control and attendance, and mobile phone unlocking. For instance, during mobile payment, the system will confirm through liveness detection that the operator is a real user rather than using the user's photo to forge the identity. With the upgrading of attack methods (such as high-fidelity 3D masks), liveness detection technology is also evolving towards deep learning-driven end-to-end models, enhancing anti-forgery capabilities by extracting more refined features (such as skin texture and blood flow signals), and has become an indispensable security barrier in facial recognition systems. Aashania Antil and Chhavi Dhiman[2] systematically sorted out the development context of face anti-deception methods, providing a clear review and reference framework for field research.

Traditional machine learning methods (such as detection based on LBP operators and color texture features) overly rely on manual feature design in specific scenarios. They have weak generalization capabilities in cross-domain scenarios involving different devices, lighting conditions, and attack types, and are difficult to adapt to a variety of deceptive tactics. For instance, Maatta et al[3]. adopted the multi-scale LBP (Local Binary Pattern) operator to achieve face spoofing detection in a single image. However, it only conducts face spoofing detection for a single image and has weak generalization ability in complex scenes (such as illumination changes and pose shifts), making it difficult to cope with diverse spoofing attack methods. Boulkenafet et al[4]. proposed a

[1]School of Computer Science and Technology, Taiyuan Normal University, Jinzhong 030619, China. [2]School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China. [3]Shanxi Key Laboratory of Intelligent Optimization Computing and Blockchain Technology, Jinzhong 030619, Jinzhong, China. [4]Research Center for the Development of Education among Ethnic Minorities in Northwest China, Northwest Normal University, Lanzhou 730070, China. [5]Department of Information Technology, Shanxi Professional College of Finance, Taiyuan 030008, China. [6]School of Information, Shanxi University of Finance and Economics, Taiyuan 030006, China. [7]Hui Qi, Rui Han, Kaige Duan, Ying Shi, Xiaobo Qi, Chenghai Gao and Lifang Ren contributed equally to this work. ✉email: gaochenghai@163.com

detection method based on color texture features, which integrates the texture features of the YCbCr and HSV color Spaces. However, it overly relies on the texture information of specific color Spaces. When facing spoofing attacks from non-target color Spaces or those with indistinct texture features, the detection performance will significantly decline. Li et al[5]. for the first time applied remote photoplethysmography (rPPG) to face liveness detection. By extracting heart rate features, they identified real faces and further distinguished real liveness from video attacks by combining LBP color and texture features. However, the remote light volume tracing method is susceptible to interference from factors such as ambient light and shooting distance, resulting in unstable extraction of heart rate features and thereby affecting the discrimination accuracy between real faces and video attacks.

Researchers have successfully introduced a series of advanced technologies into the field of face liveness detection for deep learning methods. In 2014, Yang et al.[6]. were the first to propose a face liveness detection method based on convolutional neural networks (CNN), and constructed a 13-layer neural network for extracting features from RGB images and training models. However, relying solely on RGB images for feature extraction without considering multimodal information such as depth and texture, the ability to distinguish forged faces is limited in scenarios with drastic changes in lighting or complex backgrounds. In recent years, George et al.[7]. used Vision Transformers (ViT) as the pre-trained model and achieved remarkable results in the task of face liveness detection through transfer learning, but it overly relied on the initial feature distribution of the pre-trained model. When confronted with new types of spoofing attacks that differ significantly from the pre-trained data (such as 3D printed masks), the generalization performance is prone to decline. Yu et al.[8]. proposed the TransRPPG method, which integrates the remote photoplethymmetric signal into the feature learning process and uses ViT for classification. However, due to its reliance on the remote photoplethymmetric signal, this signal is susceptible to interference from shooting distance and ambient light, resulting in insufficient stability of feature extraction in uncontrolled acquisition scenes and affecting classification accuracy. Qiao et al.[9]. successfully transformed the self-attention mechanism of the Transformer model into a convolution operation, which reduced the computational cost and improved the model performance. However, the local perception characteristics of convolution may lose the global live feature association of the face and have a weak detection ability for global forgery patterns (such as overall facial texture tampering). To further enhance the Generalization ability of the model, Shao et al.[10]. introduced the Domain Generalization technology into the field of face liviality detection for the first time. However, they did not consider optimizing the domain adaptation strategy for the task characteristics of face liviality detection. When the data distribution of multi-source domains varies greatly, it is difficult to balance the detection performance of each domain. It is prone to the problem of low accuracy in some domains. Perez et al.[11]. regarded face liveness detection as an anomaly detection problem and used the triplet contrastive loss to restrain the features learned by the model to ensure compact features of the same class and dispersed features of different classes. However, it has a high requirement for the diversity of anomaly samples (forged faces). When the abnormal sample types in the training set are not fully covered, the detection effect on unseen spoofing attacks is not good. Jourabloo et al.[12]. dealt with forged faces from the perspective of image denoising, treating forged faces as noisy samples. However, they failed to distinguish between "forged noise" and the natural noise of real images. The denoising process might mistakenly delete the living feature details of real faces, increasing the risk of misjudging real faces. Jia et al.[13]. proposed a single-domain generalization algorithm for face liveness detection. This method effectively improves the clustering effect of real faces in the shared feature space and enhances the discrimination of forged faces between different domains. However, the feature discrimination of forged faces only relies on inter-domain differences. When the forgery patterns in different domains are similar, it is difficult to effectively distinguish cross-domain forged samples. Limited generalization ability; Wang et al.[14]. proposed a novel method that decomposes feature representations into content features and style features, and uses an unordered style recombination network to extract and recombine different content and style features from the stylized feature space. However, this process may disrupt the correlation between content features and live information, resulting in the loss of key facial live details (such as micro-eye movements). It affects the accuracy of the detection. In 2023, Lin et al.[15]. proposed the DEFAEK fast learning method. This method rapidly ADAPTS to new tasks through an optimized meta-learning paradigm, enhancing the adaptability and efficiency of face anti-deception. However, the meta-learning paradigm has high quality requirements for the initial task dataset. When the initial dataset contains annotation noise or sample bias, The accuracy of the model after rapid adaptation is easily affected. The four articles published by Sareer Ul Amin et al.[16–19]. all focus on techniques such as the attention mechanism and feature fusion in deep learning, and have significant advantages in fields such as anomaly detection of surveillance videos and detection of pine wilt disease: Generally, the detection accuracy and model generalization ability have been effectively improved by introducing innovative attention modules (such as external attention, CBAM) or optimizing feature extraction methods (such as the fusion of 3D convolution and LSTM, and synthetic data supplementation). However, the coverage of some research datasets is limited, which may affect the applicability of the model in a wider range of real environments. Vitor Luiz et al.[20]. effectively captured the spatio-temporal features of human faces through residual spatio-temporal convolutional networks, enhancing the detection ability of dynamic spoofing attacks. However, this method has a weak ability to distinguish static high-simulation spoofing (such as 3D printed masks), and its real-time performance is easily affected when the model complexity is high. Antil et al.[21]. utilized the shared layer Transformer to achieve multimodal feature fusion, enhancing cross-scenario generalization capabilities and adapting to various types of deception. However, they overly relied on multimodal data input. When some modal information is missing (such as depth data being unavailable), the detection performance may significantly decline.

Therefore, in response to the above problems, especially the insufficient ability of deep feature representation, the difficulty in effectively preserving key facial details, and the difficulty in efficiently extracting live face features between different source domains, this paper proposes a high-performance adaptive fusion network

algorithm for live face detection (DBC_ResNet18). This algorithm takes ResNet-18 as the backbone network and innovatively introduces the bottleneck attention module and dynamic convolution technology, aiming to enhance the extraction ability of key facial information and the expression ability of cross-domain live features. Meanwhile, the algorithm also integrates a central difference convolution module and an adaptive instance normalization layer, combining global and local facial information to reduce detail loss, thereby enhancing the style transfer effect and model generalization ability. Ultimately, this model can accurately detect real living organisms while significantly reducing generalization errors.

The main research contributions of this article are:

- This paper takes ResNet-18 as the backbone network and innovatively introduces the bottleneck attention module and dynamic convolution technology. Among them, the bottleneck attention module can adaptively focus on key facial areas such as the eyes and nose, and the dynamic convolution can adjust the convolution kernel parameters according to the input features. The synergistic effect of the two significantly enhances the extraction accuracy of key facial information. At the same time, it effectively enhances the expression ability of live features in cross-domain scenarios, laying a high-quality foundation for subsequent feature processing.
- By introducing the central difference convolution module and the adaptive instance normalization layer, this paper constructs a more complete feature processing mechanism - the central difference convolution can enhance the detail representation ability of features and reduce the loss of edge information, while the adaptive instance normalization layer can dynamically adjust the feature distribution to adapt to data in different domains. The combination of the two, which combines global and local facial information, Not only has the feature consistency during the style transfer process been optimized, but also the generalization ability of the model in complex cross-domain scenarios such as cross-device and cross-lighting has been significantly enhanced.
- The DBC_ResNet18 model constructed based on the above technological innovations, while ensuring the core function of accurately detecting real living organisms, effectively reduces the generalization error through multi-module collaborative optimization. Experimental verification shows that this model has demonstrated excellent performance on multiple authoritative benchmark datasets such as OULU-NPU and MSU-MFSD. It takes into account both detection accuracy and efficiency, providing a technical solution that is both practical and innovative for cross-domain face liveness detection tasks.

Related theories

Facial ground marks

Facial ground marks are a set of predefined points that mark the positions and contours of the main features of the human face and are an important part of face analysis preprocessing in face anti-deception detection methods. Figure 1 is such a set of facial landmark points, which are usually represented as two-dimensional coordinates relative to the entire face. In this paper, the Dlib 68-point landmark model[22] was used, and these points are located in key areas of the face, such as the corners of the mouth, the tip of the nose, and the eyes. It can be seen from the figure that each eye is marked with six dots. The specific enlarged image is shown in Fig. 2. These six points are numbered in a clockwise direction and are used to calculate the Eye Aspect Ratio (EAR). When the eyes are open, the EAR value fluctuates within a small, stable range. When the eyes are closed, the EAR quickly drops close to zero. The calculation method is shown in Formula 1.

$$EAR = \frac{\|P_2 - P_6\| + \|P_3 - P_5\|}{2\|P_1 - P_4\|} \tag{1}$$

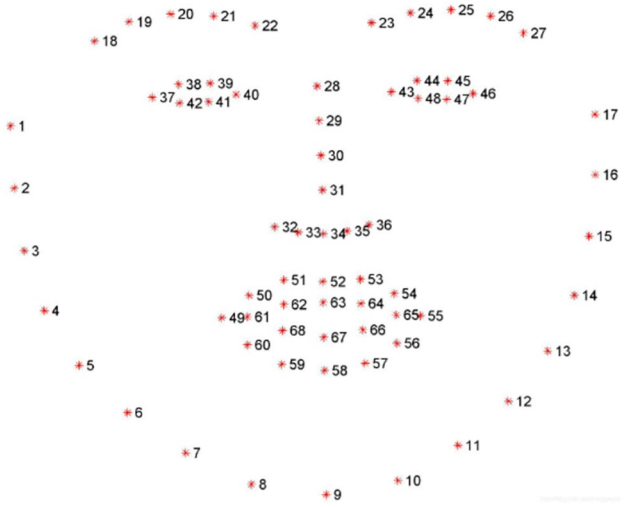


Fig. 1. Sixty-eight-point feature map of human face.

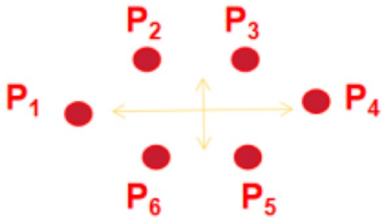


Fig. 2. Six facial landmarks of the eye.

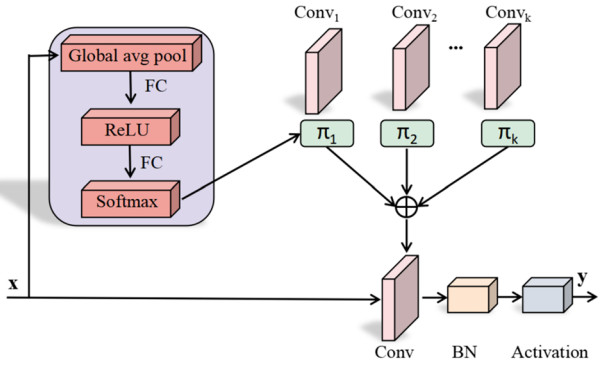


Fig. 3. Dynamic convolution structure diagram.

The study adopted the Dlib 68-point marker model, taking these predefined marker points (distributed in key facial areas such as the corners of the mouth, the tip of the nose, and the eyes, and each Eye was marked with 6 marker points numbered clockwise) as the basis, and through calculating the Eye Aspect Ratio (EAR), The aspect ratio of the eyes is used to capture the dynamic features of the eyes - when the eyes are open, the EAR fluctuates stably in a small range, and when closed, it quickly approaches zero. This feature calculation method based on facial landmark points can effectively extract detailed information related to the living body (such as eye movement). It provides important physiological feature basis for subsequent models to distinguish real faces from forged faces (such as photo and video attacks), thereby assisting in improving the accuracy and reliability of cross-domain face liveness detection.

Dynamic convolution

The traditional static convolutional network model fixes its parameters after training. Therefore, during the inference phase, regardless of input variations, the model always uses the same parameters. This approach does not account for feature differences between different input samples, which may limit the model's generalization ability and adaptability.

To better extract facial image features and enhance network performance, we introduce dynamic convolution[23] for feature extraction. Its structure is shown in Fig. 3. Unlike traditional convolution methods, dynamic convolution does not use fixed convolutional kernels at each layer. Instead, it dynamically adjusts the weights of the convolutional kernels based on the features of the input data. Specifically, each convolutional kernel consists of k sub-kernels. The input data extracts the weights of these k convolutional kernels through an attention mechanism, and then aggregates them using these weights $\pi_k(x)$. The process of calculating $\pi_k(x)$ includes averaging pooling of global features to extract global spatial information. This is followed by transforming it into k dimensions through two fully connected layers, and finally normalizing the results. The calculation formula for the attention weights is as follows:

$$\begin{aligned} y &= g\left(\tilde{W}^{\mathrm{T}}(x)x + \tilde{b}(x)\right) \\ \tilde{W}(x) &= \sum_{k=1}^{K} \pi_k(x) \cdot \tilde{W}_k, \tilde{b}(x) = \sum_{k=1}^{K} \pi_k(x)\tilde{b} \\ 0 &\le \pi_k(x) \le 1, \sum_{k=1}^{K} \pi_k(x) = 1 \end{aligned} \tag{2}$$

Here, x is the input data; y is the output result; g is the activation function; W is the weight matrix; b is the bias term; and π_k is the attention coefficient of the $\tilde{W}_k^T(x) + \tilde{b}_k$ linear function. Aggregate weights $\tilde{W}(x)$ and bias $\tilde{b}(x)$ are functions of the input. They maintain the same attention mechanism.

This paper uses dynamic convolution to break through the limitation of fixed parameters after training in traditional static convolution. It can dynamically adjust the weights of the convolution kernel according to the feature differences of the input face image (such as differences in illumination, posture, and types of spoofing attacks in different domains), and work in synergy with the bottleneck attention module in the content feature extraction stage. It not only enhances the precise capture of key facial features such as the eyes and nose, but also improves the model's adaptability to cross-domain facial features, effectively alleviating the insufficient generalization ability caused by fixed parameters in traditional convolution. This lays a foundation for the subsequent fusion of style and content features and the final improvement of liveness detection accuracy.

Bottleneck attention module

The attention mechanism simulates human visual or cognitive processes. It assigns different weights to various input elements or features to focus on and process important information. In deep learning, the attention mechanism can be applied to neural network structures to enhance model expressiveness and performance.

In this paper, we introduce a mixed attention mechanism that combines channel and spatial attention-the Bottleneck Attention Module[24]. Its network structure is illustrated in Fig. 4. The upper branch implements the channel attention mechanism. First, we apply global average pooling to the input feature map to generate a $1 \times 1 \times C$ feature vector. This vector is fed into a fully connected layer to reduce the number of channels, followed by a ReLU activation function to introduce non-linearity. Next, it passes through another fully connected layer to restore the channel count, and we use the Sigmoid activation function to generate the channel attention weight vector. This attention weight vector is applied to the input feature map, weighting each channel to produce the adjusted feature map $\mathrm{M}_c(\mathrm{F})$. Then, the spatial attention branch reduces the channel count using a 1×1 convolution. It then fuses features using two 3×3 dilated convolutions and compresses the channel count with another 1×1 convolution. Finally, the feature maps generated by channel attention $\mathrm{M}_c(\mathrm{F})$ and spatial attention $\mathrm{M}_s(\mathrm{F})$ are fused together. This fused feature map is connected to the original input via a skip connection, ensuring that the output feature map maintains the same dimensions.

In the Bottleneck Attention Module module, in the channel attention branch of this paper, after global average pooling, a $1 \times 1 \times C$ feature vector is generated (C is the number of channels of the input feature map, which needs to match the number of output channels of the corresponding layer of the backbone network ResNet-18). Typically, the number of output channels at each stage of ResNet-18 is 64, 128, 256, and 512. Here, C is set to 256 to meet the requirements of feature extraction. The channel compression ratio of the fully connected layer is 1/4 (for example, compressing 256 channels to 64 channels), and then restoring to the original number of channels. The spatial attention branch adopts 1×1 convolution (the number of convolution kernels is usually half of the number of input channels), two 3×3 dilated convolution (with the dilation rate set to 2 to expand the receptive field), and 1×1 convolution (the number of convolution kernels is consistent with the number of output channels of channel attention).

By using this module (through a hybrid design of channels and spatial attention), not only can the importance of channel features in key facial areas such as the eyes and nose be highlighted, but also the spatial receptive field can be expanded to capture local details and global correlations. Combined with dynamic convolution, it enhances the ability to extract key features of cross-domain faces. Meanwhile, the skip connection avoids feature degradation. It provides high-quality feature support for the subsequent decomposition of content-style features and the improvement of model generalization capabilities.

Adaptive instance normalization algorithm

Adaptive Instance Normalization[25] is used for image style transfer. It adjusts the mean and variance of image features to achieve style changes while preserving the content information of the input image and blending it with the target style. This method has a fast inference speed, making it suitable for real-time applications. To combine content features fc and style features fs, this paper establishes a style reconstruction layer using Adaptive Instance Normalization layers, convolution operators, and residual mappings. Specifically, after the content feature x undergoes instance normalization, it is adjusted using the parameters γ and β generated from

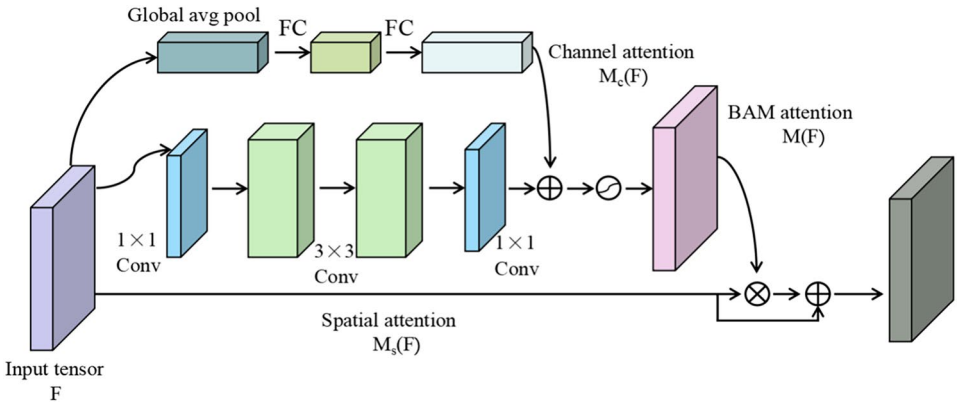


Fig. 4. Bottleneck attention module structure diagram.

the style input y. This adjustment makes the statistical properties of the normalized content feature x closer to those of the style feature y. The calculation formula is as follows:

$$\mathrm{AdaIN}(x, y, \beta) = y(\frac{x - \mu(x)}{\sigma(x)}) + \beta \tag{3}$$

In this context, $\mu(\cdot)$ and $\sigma(\cdot)$ represent the channel mean and standard deviation, respectively. The calculation of the affine parameters γ and β is as follows:

$$\begin{aligned} \gamma, \beta &= \mathrm{MLP}[\mathrm{GAP}(f_s)] \\ z &= \mathrm{ReLU}\left[\mathrm{AdaIN}\left(K_1 \otimes f_c, \gamma, \beta\right)\right] \\ \mathrm{SAL}(f_c, f_s) &= \mathrm{AdaIN}(K_2 \otimes z, \gamma, \beta) + f_c \end{aligned} \tag{4}$$

In this context, K_1 and K_2 are three 3×3 convolution kernels. The symbol $\otimes$ denotes the convolution operation, and z is an intermediate variable. To integrate content and style features, the output of the Adaptive Instance Normalization layer is added to the original feature *fc*, resulting in the fused feature (*fc*,*fs*).

In this paper, the Adaptive instance normalization algorithm is used. On the one hand, it works in synergy with Central Difference Convolution. While performing statistical normalization on the feature map, more useful feature information is retained, which solves the problem that the traditional AdaIN layer may lose fine structures and differences when processing the feature map, resulting in insufficient feature expression learned by the model, and enhances the model's perception ability of facial image details and the quality of feature extraction. On the other hand, by dynamically adjusting the feature distribution to adapt to data from different domains and combining global and local face information to optimize the feature consistency during the style transfer process, the generalization ability of the DBC_ResNet18 model in complex cross-domain scenarios such as cross-device and cross-lighting has been significantly enhanced.

Central difference convolution

The center differential convolution[26,27] is an optimization scheme proposed by Yu et al. It is mainly applied in the task of face liveness detection. Center differential convolution enhances the aggregation of image features and gradient information. This allows it to capture details in facial images more effectively. Unlike traditional convolution, which performs a weighted sum of pixel values within the receptive field and scans the image row by row, center differential convolution first processes the pixel values in the receptive field with respect to the center pixel value. This generates new features before the convolution operation. The specific structure is shown in Fig. 5. The convolution operation formula is as follows:

$$y = \theta \cdot \sum_{p_n \in \mathbb{R}} w(p_n) \cdot (x(p_0 + p_n) - x(p_0)) + (1 - \theta) \cdot \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n) \tag{5}$$

Here, $w(p_n)$ is the weight of the convolution operator at the point p_n. x represents the input feature map in the convolution operation. $x\ (p_0 + p_n)$ is the gray value of the feature map at the point $p_0 + p_n$.

In our model, center differential convolution does not completely eliminate traditional convolution. Instead, it introduces a balance parameter, θ, to adjust the contributions of both methods, θ is set to 0.7.

In this paper, central difference convolution is used, which can enhance the aggregation ability of facial image features and gradient information, capture facial details (such as edge and texture information of key areas) more effectively, and make up for the shortcomings of traditional convolution that only weights and sums pixel values within the receptive field and is prone to losing subtle structural differences. Meanwhile, it is co-optimized with Adaptive Instance Normalization to retain more useful feature information when performing statistical normalization on the feature map, avoiding the problem of important information loss that may occur when

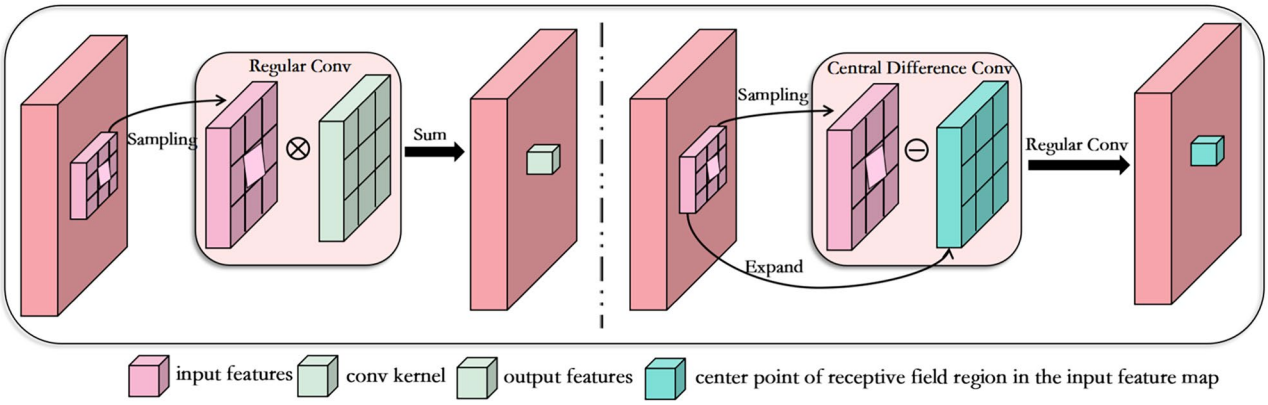


Fig. 5. Standard convolution and central differential convolution.

traditional AdaIN layers process the feature map. The feature expression ability and detail perception ability of the model have been significantly enhanced.

Contrastive learning of stylized features

In cross-domain scenarios, specific domain style features may interfere with the judgment of facial liveness. To address this issue, this paper employs contrastive learning to highlight liveness-related style features and suppress domain influence. We input the original feature $S(x_i, x_i)$ into the classifier and use a binary classification loss function L_{cls} for supervised training. For the style-reorganized feature $S(x_i, x_i^*)$, we measure its difference from the original feature using cosine similarity.

$$Sim(a, b) = -\frac{a}{\|a\|_2} \cdot \frac{b}{\|b\|_2} \tag{6}$$

Here, a and b represent the two compared features. In contrastive learning, we first use the original feature $S(x_i, x_i)$ as the reference point in the stylized feature space and fix its position. Then, we use the style-reorganized feature $S(x_i, x_i^*)$ to adjust its proximity to the original feature based on liveness information. During this process, we perform backpropagation using shuffled assembled features without changing the position of the original feature. This integration enriches the stylized features with valuable liveness information.

$$L_{contra} = \sum_{i=1}^{N} E_q(x_i, x_i^*) \cdot Sim(\text{stopgrad}(a), b) \tag{7}$$

Here, $a = S(x_i, x_i)$ and $b = S(x_i, x_i^*)$. The term $\text{Eq}(x_i, x_i^*)$ indicates the consistency of liveness labels between x_i and x_i^*.

$$E_q(x_i, x_i^*) = \begin{cases} +1, \text{label}(x_i) == \text{label}(x_i^*), \\ -1, \text{otherwise} \end{cases} \tag{8}$$

This paper uses contrastive learning of stylized features, which can effectively highlight the style features related to liveness in cross-domain scenarios and suppress the interference caused by domain differences. By deeply integrating stylized features with liveness information, the detailed information key to liveness judgment in the features is enriched. Assist in enhancing the model's ability to distinguish cross-domain forged faces (such as attacks on photos or videos from different devices or under different lighting conditions); Meanwhile, this method collaborates with techniques such as content-style feature separation and adversarial training, further enhancing the feature expression ability and cross-domain generalization ability of the DBC_ResNet18 model.

Domain adversarial learner

Existing methods for face anti-spoofing domain generalization mainly enhance the model's generalization ability by extracting shared features. However, constructing a unified feature space becomes complex due to the distribution differences of spoofed faces across different domains. To address this, this paper reduces the differences between real face features from different source domains using adversarial learning methods to improve model generalization. Specifically, after extracting content features, the features are sent to a domain discriminator. This discriminator is used to distinguish the source of the features. During training, we adjust the parameters of the feature extractor by optimizing the loss of the domain discriminator.

To reduce training complexity, we introduce a gradient reversal layer[28], which adjusts the gradient by multiplying it with a negative scalar during backpropagation. For multiple source domains, we use standard cross-entropy loss to optimize the adversarial learning network. The specific calculations are detailed in Equation 9.

$$\min_D \max_G L_{adv}(G, D) = -E_{(x,y)\sim(X,Y_D)} \sum_{i=1}^{M} 1[i = y] \log D(G(x)) \tag{9}$$

This paper uses this module to break through the limitation of existing cross-domain face anti-deception methods that only rely on extracting shared features to improve generalization ability. Through adversarial learning, it reduces the distribution differences of real face features in different source domains, effectively alleviating the model adaptation problem caused by the heterogeneity of multi-source domain data distribution.

Loss function

This paper defines an overall loss function for network training by combining classification loss, adversarial loss, and contrastive loss. The specific expression is as follows:

Here, L_{overall}, L_{cls}, L_{adv}, and L_{contra} represent the overall loss, classification loss, adversarial loss, and contrastive loss, respectively. λ_1 and λ_2 are two hyperparameters used to adjust the weights between the different loss functions. This paper employs an end-to-end training approach to generate a more universal domain-shared feature space. This enhances the model's adaptability to the target domain.

$$L_{\text{overall}} = L_{cls} + \lambda_1 \cdot L_{adv} + \lambda_2 \cdot L_{contra} \tag{10}$$

This paper achieves multi-dimensional performance improvement of the model through multi-loss collaborative optimization: The classification loss ensures that the model has the basic ability to classify real and fake faces, providing a core discrimination basis for liveness detection tasks; Adversarial loss reduces the distribution differences of features in different source domains through domain adversarial learning, effectively alleviates the problem of cross-domain data heterogeneity, and enhances the generalization ability of the model in complex scenarios such as cross-device and cross-lighting. Contrast loss enhances the effectiveness of feature expression by strengthening the association between stylized features and live information, highlighting key live features and suppressing domain interference. It provides an efficient training method for cross-domain face anti-spoofing detection that takes into account classification accuracy, cross-domain generalization and feature discrimination.

DBC_ResNet18 face anti-spoofing detection algorithm

This article presents an improved SSAN_R[14] face liveness detection model-DBC_ResNet18. The overall framework of the algorithm includes three modules: data preprocessing, feature extraction, and liveness detection, as shown in Figure 6. It mainly covers the following content:

- Data preprocessing. All sample image data is used, while video data extracts one frame every 10 seconds. After processing the images, MTCNN[29] is used for face detection. The detected faces are cropped and resized to 224×224 pixels for RGB input. Additionally, we use a dense face alignment method (PRNet[30]) to generate a 32×32 depth map for real faces. For fake faces, the depth map is all zeros.
- Feature extraction. In the ResNet-18 feature extractor, a bottleneck attention module is introduced to enhance the model's focus on key features. Dynamic convolution is used to optimize the capture of spatial features. Then, the content and style feature extractors are used to separately extract content and style features. We combine central difference convolution with adaptive instance normalization to improve feature representation and achieve style reorganization. Through adversarial learning, the differences in face features from different source domains are reduced, making the model more stable when handling faces from various sources.
- Face anti-spoofing detection. The liveness detection module performs binary classification on the input face images to determine their authenticity. If the detection threshold is greater than or equal to 0.9, it is classified as a real face. If it is less than 0.9, it is classified as a fake face.

The overall network structure of the DBC_ResNet18 model is shown in Fig. 7. The upper part of Fig. 7 is the visualization diagram of the improved DBC_ResNet18 model based on the original SSAN-R model. On the basis of the original SSAN-R network structure, new elements such as dynamic convolution, bottleneck concern module, and central difference convolution have been added. This figure presents the overall architecture of the DBC_ResNet18 face anti-deception detection algorithm. Firstly, the ResNet-18 Feature Extractor extracts common features from the input face images in multiple domains (domain 1 to domain N), and then it is divided into two paths. The Content features are extracted by the Content Feature Extractor one way, and then input into the Discriminator through the Gradient Inversion Layer (GRL) to achieve domain adversarial learning. Another way acquires Style features through Style Feature Extractor, and both are input into the Style Assembly module together. Meanwhile, the underlying network enhances feature expression by integrating DynamicConv (dynamic convolution) and BAM (Bottleneck Attention Model) through basic operations such as 7×7 convolution and MaxPooling, and optimizes style reconstruction by using CDC_AdaIN (central difference convolution combined with adaptive instance normalization). Finally, through Avg Pooling and FC (Fully connected Layer) output, combined with the classification loss and contrast loss of Classifier, cross-domain

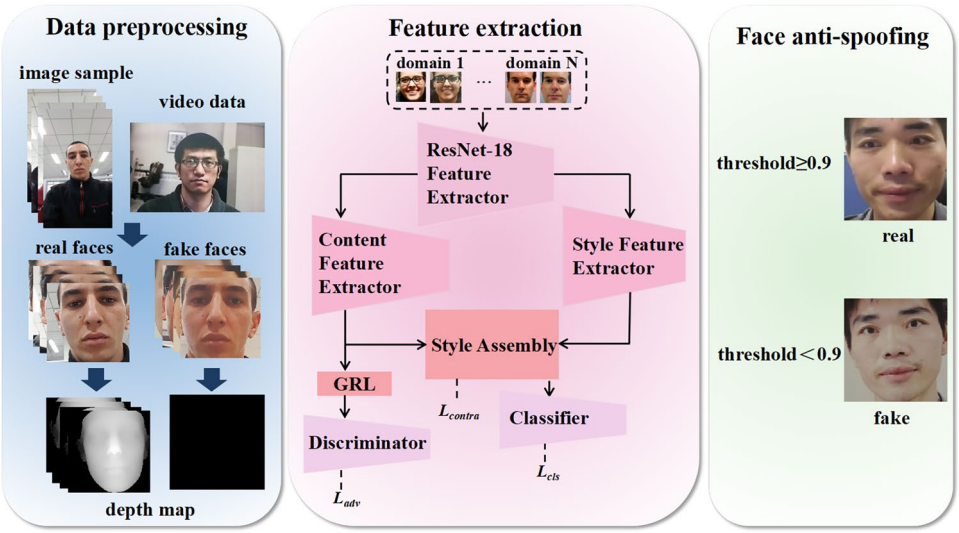


Fig. 6. Overall framework diagram.

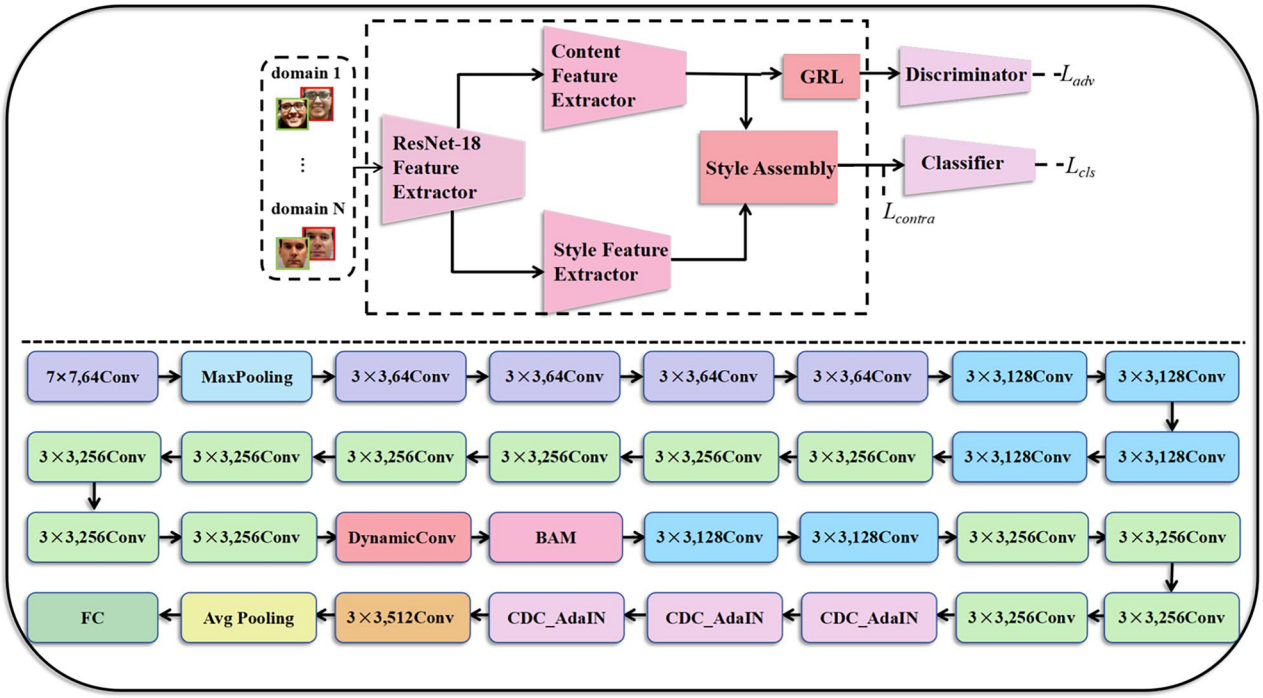


Fig. 7. DBC_ResNet18 model architecture diagram.

face authenticity detection is achieved collaboratively. The lower part of Fig. 7 is the detailed network structure diagram of the entire DBC_ResNet18 model.

The Feature Generator first extracts the common features of images in different domains based on ResNet-18, maps them to the shared feature space, and then inputs the obtained common feature vectors into the content feature extractor and the style feature extractor respectively, decomposing them into content features *fc* and style features *fs*. Among them, the content feature extraction process integrates Dynamic convolution (dynamic aggregation of convolution kernel weights to adapt to input differences) and Bottleneck attention model (fusing channels and spatial attention to focus on the key areas of the face). Strengthen the ability to extract key features and obtain domain-invariant representations. Take domain-invariant representations to lay the foundation for subsequent feature processing.

Style reconstruction builds the style reconstruction layer by means of Adaptive Instance normalization (AdaIN) and central difference convolution. First, instance normalization is performed on the content features, and then the scaling factor and offset factor generated by the style input are used to adjust their statistical characteristics to achieve the adaptation of content and style features. At the same time, by combining central difference convolution (introducing balanced difference), more useful feature information is retained during statistical normalization, avoiding the loss of fine structures in traditional AdaIN layers and improving the quality of feature expression.

Contrastive learning of stylized features takes the original stylized features as the fixed reference point, calculates the difference between them and the stylized features using cosine similarity, and adjusts the feature proximity based on the consistency of the living labels of the two through the contrastive loss function. Without changing the position of the original features, By using the backpropagation of scrambled assembly features, the live information in the stylized features is enriched and domain interference is suppressed.

The domain adversarial learner inputs the content features into the domain discriminator to distinguish the feature source domain. The Gradient Inversion Layer (GRL) is introduced during training to optimize the adversarial learning network through standard cross-entropy loss. By minimizing the discriminator loss and maximizing the feature extractor loss, the distribution differences of real facial features in different source domains are reduced, the problem of data heterogeneity in multiple source domains is alleviated, and the cross-domain generalization ability of the model is improved.

The Loss function integrates classification loss, adversarial loss and contrastive loss through the total loss function. Among them, the classification loss guarantees the basic binary classification ability of the model, the adversarial loss optimizes the alignment of domain features, and the contrastive loss strengthens the association between style features and live information. Through end-to-end training, a universal domain shared feature space is constructed to enhance the adaptability of the model to the target domain.

The DBC_ResNet18 face anti-deception detection algorithm integrates the above modules and is based on the SSAN-R architecture. Firstly, it obtains 224×224 pixel RGB images and 32×32 depth maps through data preprocessing, and then extracts and decomposes the features through the Feature Generator. By means of Style reconstruction to optimize the feature representation and using Contrastive learning of stylized features to strengthen the living-related features The Domain differences are reduced through the Domain adversarial

Parameters	Values
Model type	Sequential
Learning rate	0.0001, 0.01
Learning rate schedule	Cosine decay, Warmup
Batch size	16
Epoch	60
Optimizer	Adam, SGD
Gamma	0.5
Activation layer	ReLU
Kernel size	3×3, 7×7

Table 1. Model parameters.

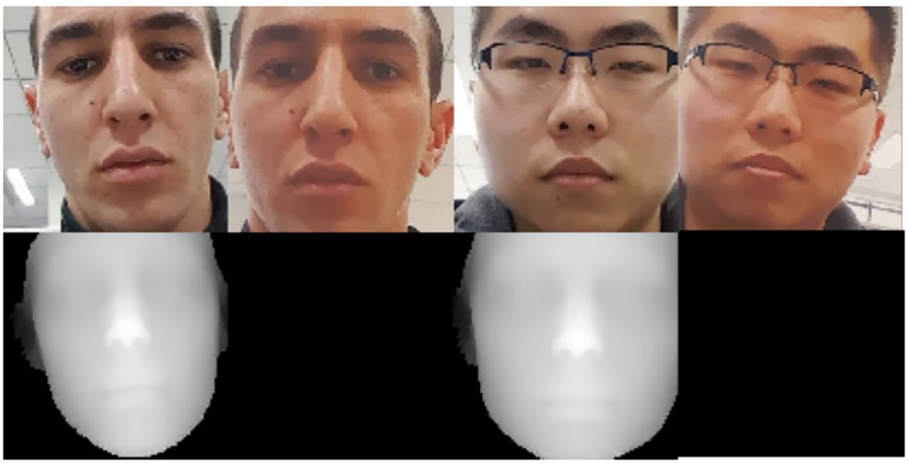


Fig. 8. Examples from the OULU-NPU dataset.

learner. Finally, the model is trained with the Loss function of multi-loss fusion to achieve binary classification of real/fake faces (detection threshold 0.9). It performed outstandingly on four benchmark datasets such as OULU-NPU and MSU-MFSD. Balancing detection accuracy and efficiency (single graph prediction time 0.0388 seconds). Provide effective technical solutions for cross-domain face anti-deception.

Parameter settings

The experiments were conducted on a 64-bit Ubuntu 7.5.0-3ubuntu1-18.04 platform. The CPU frequency is 2.40 GHz, and the memory is 251 GB. The improved SSAN_R network was implemented using Python and the PyTorch framework. During training, the hyperparameters λ_1 and λ_2 were both set to 1.For experiments on the O&C &I to M dataset, the Adam optimizer was used with a learning rate of 0.0001 and weight decay of 0.00005. In the experiments with the other three dataset protocols, the SGD optimizer was employed. The initial learning rate was set to 0.01, with a momentum of 0.9 and a weight decay of 0.0005. The learning rate was reduced by half every ten epochs until the 60th epoch.The specific model parameter settings are shown in Table 1.

Dataset

This paper uses four public datasets for experimental validation. They are OULU-NPU[31] (abbreviated as O), MSU-MFSD[32] (abbreviated as M), CASIA-FASD[33] (abbreviated as C), and Replay-Attack[34] (abbreviated as I).

(1) OULU-NPU dataset

The OULU-NPU dataset contains 4,950 video clips. The frame rate is 30 Hz, and the resolution is 1920×1080 pixels. It includes both real and fake videos. The data was collected by 55 volunteers in three different scenarios using six types of smartphones. The ratio of real faces to fraudulent faces is 1:4. As shown in Fig. 8, the upper half of the images displays real and fake faces. The lower half shows the corresponding facial depth maps. The depth map of attack faces appears almost completely black, with pixel values close to zero due to their flat structure. In contrast, the real faces show significant depth differences in key areas, such as the mouth, nose, and forehead.

(2) MSU-MFSD dataset

The MSU-MFSD dataset contains 440 video clips provided by 55 subjects. The videos were recorded using the built-in camera of a MacBook Air 13 and the front camera of a Google Nexus 5 phone. The dataset includes real videos from both computers and phones. The fake videos consist of three types: high-resolution video playback, mobile video playback, and print attacks. As shown in Fig. 9, the upper half displays real faces and fake faces. The lower half shows the corresponding facial depth maps.

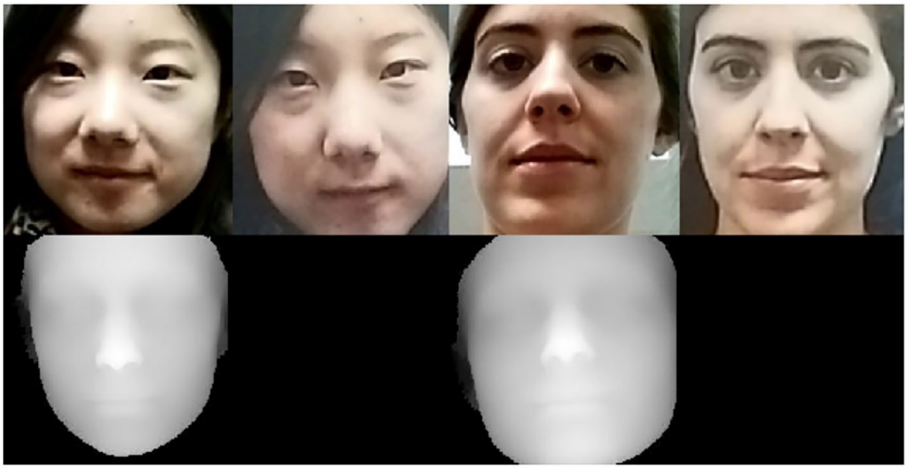


Fig. 9. Examples from the MSU-MFSD dataset.

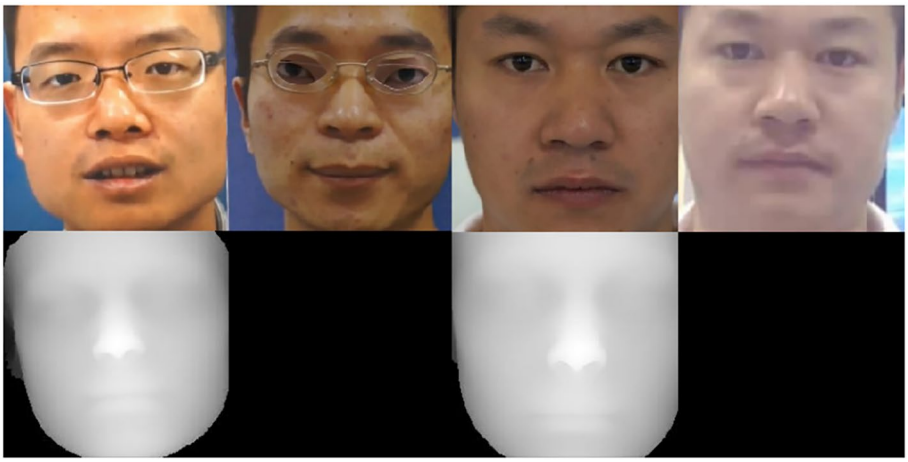


Fig. 10. Examples from the CASIA-FASD dataset.

(3) CASIA-FASD dataset

The CASIA-FASD dataset contains 600 video clips from 50 different subjects. Among these, 150 are real face videos, and 450 are from spoofing attacks. The spoofing attacks include distorted photo attacks, cut photo attacks, and video attacks. As shown in Fig. 10, the upper half displays real faces and fake faces. The lower half shows the corresponding facial depth maps.

(4) Replay-attack dataset

The Replay-Attack dataset was released by IDIAP in 2012. It contains 1,300 video clips from 50 individuals. The dataset includes 300 real face videos and 1,000 videos covering various spoofing attacks. These videos were recorded under different lighting conditions.As shown in Fig. 11, the upper half displays real faces and fake faces. The lower half shows the corresponding facial depth maps.

Evaluation metrics

This paper uses the Half Total Error Rate and Area Under Curve to evaluate algorithm performance. The task of face liveness detection is a binary classification problem. The evaluation metrics are calculated based on the confusion matrix. The confusion matrix compares the actual labels with the predicted labels. It shows the classification results for each category. Positive samples are real faces, while negative samples are spoofed faces. The data is divided into true positive, false negative, false positive, and true negative samples, as detailed in Table 2.

Here, TP represents the number of true positive samples accurately identified, FN is the number of positive samples incorrectly labeled as negative, FP is the number of negative samples incorrectly classified as positive, and TN is the number of true negatives correctly classified. Based on this matrix, the following evaluation metrics are defined:

(1) False acceptance rate

Fig. 11. Examples from the Replay-Attack dataset.

Actual class	Predicted class	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Table 2. Confusion matrix.

The false acceptance rate refers to the proportion of spoofed faces incorrectly recognized as genuine faces, calculated using the method shown in Eq. (11).

$$FAR = \frac{FP}{FP + TN} \tag{11}$$

(2) False reject rate

The false reject rate represents the proportion of genuine faces incorrectly classified as spoofed faces, calculated using the method shown in Eq. (12).

$$FRR = \frac{FN}{FN + TP} \tag{12}$$

(3) Half total error rate

The half total error rate is the average of the false acceptance rate and the false reject rate, calculated as shown in Eq. (13). A lower HTER value indicates better model performance.

$$HTER = \frac{FAR + FRR}{2} \tag{13}$$

(4) True positive rate

The true positive rate measures the proportion of positive samples that are correctly classified as positive. The calculation formula is as follows:

$$TPR = \frac{TP}{TP + FN} \tag{14}$$

(5) False positive rate

The false positive rate measures the proportion of negative samples that are incorrectly predicted as positive. The calculation formula is as follows:

$$FPR = \frac{FP}{FP + TN} \tag{15}$$

The ROC curve is used to evaluate the performance of classification models. It does this by calculating the true positive rate and false positive rate at different thresholds and plotting these points to form a curve in a coordinate system. The area under the curve indicates the region enclosed by the ROC curve and the coordinate axes, with values typically ranging from 0.5 to 1.0, used to measure the overall performance of the model.

In addition, we also assess the algorithm's performance through the following metrics: the number of model parameters, the number of floating-point operations per second, and the prediction time for a single image. The number of parameters reflects the actual storage space occupied by the model, FLOPs measures the computational complexity of the model, while the prediction time for a single image demonstrates the model's real-time performance.

Experimental results and analysis

Experimental results

Cross-dataset experiments

To evaluate the generalization ability of our method across different datasets, we conducted cross-dataset experiments. The selected four datasets have different attack methods, device types, lighting conditions, background settings, and crowd characteristics. Therefore, there are significant differences among them.In the experiments, we used a leave-one-out method. One dataset was used as the test target, while the other three served as training sources. Specifically, we carried out four tasks: Using OULU-NPU, CASIA-FASD, and Replay-Attack as the source domains to test on MSU-MFSD (O&C &I to M).Using OULU-NPU, MSU-MFSD, and Replay-Attack as the source domains to test on CASIA-FASD (O&M &I to C).Using OULU-NPU, CASIA-FASD, and MSU-MFSD as the source domains to test on Replay-Attack (O&C &M to I). Using MSU-MFSD, CASIA-FASD, and Replay-Attack as the source domains to test on OULU-NPU (I&C &M to O). As shown in Table 3, the proposed algorithm performed excellently across the four datasets. In the O&C &I to M experiment, the HTER and AUC results were 0.95% and 99.42%, respectively. In the I&C &M to O experiment, the HTER and AUC were 6.22% and 97.55%. Compared to the CSEFO algorithm, HTER improved by 5.38 and 4.37 percentage points, and AUC improved by 0.91 and 2.31 percentage points.In the O&M &I to C experiment, the HTER and AUC were 0.06% and 99.88%. In the O&C &M to I experiment, the HTER and AUC were 2.54% and 98.47%. Compared to the IADG algorithm, HTER improved by 8.64 and 8.08 percentage points, and AUC improved by 3.44 and 3.97 percentage points.These results indicate that using content and style feature extractors to separate features significantly enhances their expressive ability. The contrastive learning method further enriches the style features of live information. Additionally, the bottleneck attention module, dynamic convolution, and the adaptive normalization algorithm that fuses center differential convolution also effectively improve the generalization ability of the cross-scene live detection model. The integration of these techniques significantly enhances the model's detection performance and robustness in various scenarios.

Ablation experiments

To verify the actual effects and contributions of each module in this method, four ablation factors were proposed: center differential convolution, bottleneck attention module, dynamic convolution, and the base model SSAN-R. Ablation analysis was conducted to assess the effectiveness of these factors.

1. Using SSAN-R as the base network structure.
2. Adding center differential convolution in the adaptive instance normalization layer.
3. Introducing the bottleneck attention module into the SSAN-R structure.
4. Incorporating the dynamic convolution module into the SSAN-R structure.
5. Integrating center differential convolution, bottleneck attention module, and dynamic convolution module into the SSAN-R network to form the final algorithm designed in this paper.

We validated these methods on four evaluation tasks (O&C &I to M, O&M &I to C, O&C &M to I, and I&C &M to O). The experimental results are shown in Table 4. The data in the table indicate that the proposed algorithm demonstrates significant improvements in results and generalization ability.

Method	O&C &I to M		O&M &I to C		I&C &M to O	
	HTER (%)	AUC (%)	HTER (%)	AUC (%)	HTER (%)	AUC (%)
MADDG[10]	17.69	88.06	24.50	84.51	22.19	84.99
RFM[35]	13.89	93.98	20.27	88.16	17.30	90.48
SSDG-M[13]	16.67	90.47	23.11	85.45	18.21	94.61
D2AM[36]	12.70	95.66	20.98	85.58	15.43	91.22
DRDG[37]	12.43	95.81	19.05	88.79	15.56	91.79
ANRL[38]	10.83	96.75	17.85	89.26	16.03	91.04
SSAN(R)[14]	6.67	98.75	10.00	96.67	8.88	96.79
AMEL[39]	10.23	96.62	11.88	94.39	18.60	88.79
EBDG[40]	9.56	97.17	18.34	90.01	18.69	92.28
IADG[41]	5.41	98.19	8.70	96.44	10.62	94.50
CSEFO[42]	6.33	98.51	12.05	95.44	8.38	96.88
Ours	**0.95**	**99.42**	**0.06**	**99.88**	**2.54**	**98.47**

Table 3. The comparison experiment of the proposed method and mainstream domain generalization face liveness detection algorithms on four datasets. Significant values are in bold.

Method	O&C & I to M		O & M & I to C		I & C & M to O		O & C & M to I	
	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
SSAN-R	6.67	98.75	10.00	96.67	8.88	96.97	13.72	93.63
CDC	5.03	97.71	1.08	98.89	1.69	98.35	9.54	94.58
BAM	1.02	99.07	3.51	97.97	3.89	97.82	4.41	97.84
DynamicConv	5.13	97.35	0.13	99.87	2.85	98.27	8.60	95.22
DBC_ResNet18	**0.95**	**99.42**	**0.06**	**99.88**	**2.54**	**98.47**	**6.22**	**97.55**

Table 4. Different components on the generalization performance of the proposed method. Significant values are in bold.

Method	M&I to C		M&I to O	
	HTER (%)	AUC (%)	HTER (%)	AUC (%)
MADDG[10]	41.0	64.3	39.3	65.1
SSDG-M[13]	31.9	71.3	36.0	66.9
DR-MD-Net[43]	31.7	75.2	34.0	72.7
ANRL[38]	31.1	72.1	30.7	74.1
SSAN(M)[14]	30.0	76.2	29.4	76.6
EBDG[40]	27.9	75.8	25.9	78.3
Ours	**21.05**	**86.4**	**28.5**	**77.2**

Table 5. Comparison of face anti-spoofing detection results under restricted source domain conditions.

Finite source domain experiments

To further evaluate the generalization ability of the proposed method, we limited the number of source domains. We chose the MSU-MFSD and Replay-Attack datasets, which have significantly different feature distributions, as the source domains. We used the OULU-NPU and CASIA-FASD datasets for testing. For these two test datasets, we performed the M&I to C and M&I to O tasks. The results are shown in Table 5. Despite the limited amount of source domain data, our method still achieved good results. In the M&I to C experiment, the HTER and AUC metrics reached 21.05% and 86.4%, respectively. Compared to previous methods, HTER improved by 6 percentage points, and AUC improved by 11 percentage points. In the M&I to O experiment, the HTER and AUC metrics were 28.5% and 77.2%, respectively. This further validates the good generalization ability of the proposed model. Our algorithm enhances the face liveness region using bottleneck attention and dynamic convolution. It focuses on critical facial information. Additionally, the adaptive instance normalization algorithm, which integrates center difference convolution, allows the model to align and enhance features more precisely when handling different facial characteristics. This improves overall performance.

Model complexity analysis

To study the impact of the proposed method on model complexity, we conducted experiments on the I&C &M to O protocol. We used parameters, floating-point operations per second, single-image prediction time, and half-error rate as metrics. The results are shown in Fig. 12.Here, Basemodel represents the SSAN-R network. Model One represents the SSAN-R combined with the center difference convolution network. Model Two represents the SSAN-R combined with both center difference convolution and the bottleneck attention module. Model Three represents our final model, which is the SSAN-R combined with center difference convolution, bottleneck attention, and dynamic convolution kernels.The experimental data show that Model Two achieved a significant improvement in detection accuracy with only a small increase in parameters and computational load, making it more advantageous than Model One. Furthermore, Model Three optimized performance even further by introducing dynamic convolution kernels, reducing the detection time for a single image to 0.0388 seconds.Overall, the experimental results demonstrate that the proposed method successfully enhances detection efficiency and accuracy without significantly increasing model complexity. The effectiveness of these improvements has been fully validated.

Visualization analysis

Class Activation Map is a visualization technique used to display the feature responses of deep learning models. By showing areas of high response to a specific category within an image, CAM helps to explain the model's decision-making process. It reveals which regions of the image contribute significantly to the model's classification. Using a heatmap representation, CAM indicates the importance the neural network assigns to different areas of the input image, clarifying which parts affect the final classification result. In this study, we utilized Grad-CAM[44] to perform CAM visualization on the proposed method. As shown in Fig. 13, the results of Grad-CAM illustrate the performance across three cross-dataset testing tasks: I&C &M to O, O&C &I to M, and O&M &I to C. The first, third, and fifth rows of the image present real face photos obtained from the OULU-NPU, MSU-MFSD, and CASIA-FASD datasets, respectively. Correspondingly, the second, fourth, and sixth rows

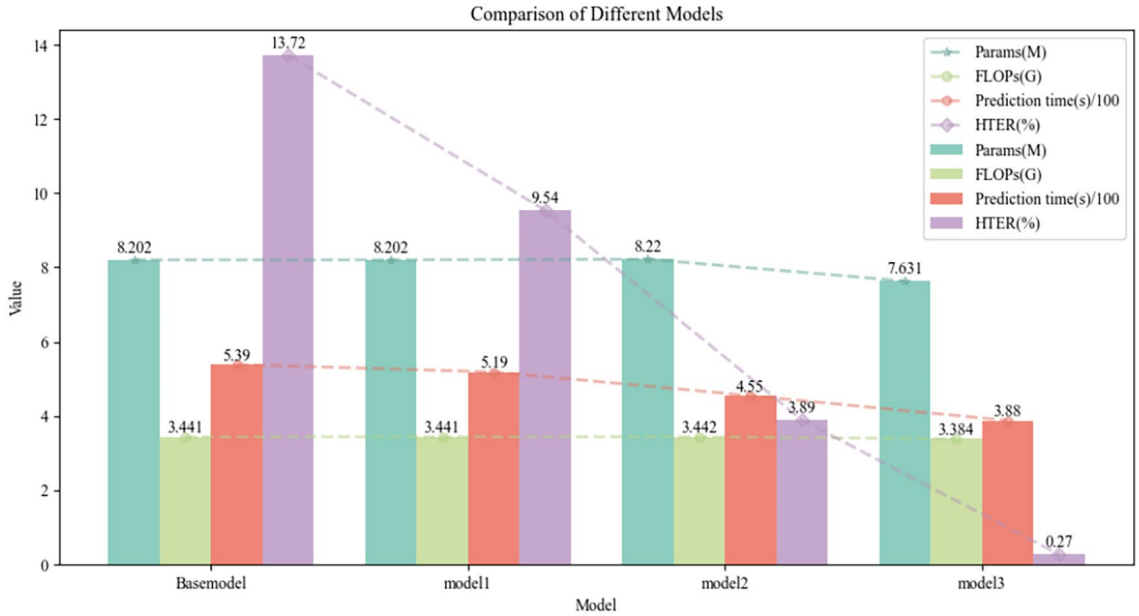


Fig. 12. Comparison of complexity among different models.

display printed face photos and facial images from video replay attacks within these datasets. The analysis results indicate that the proposed method focuses more on feature extraction from critical facial regions, such as the eyes, nose, and mouth, demonstrating good robustness under varying angles and lighting conditions.

Real-time testing results

We utilized OpenCV to access the camera and employed dlib for face detection. By integrating the proposed model with blink detection technology, we achieved real-time face recognition. The results are shown in Fig. 14, where (a) represents the recognition results for real faces, (b) illustrates the recognition results for mobile replay attack faces, and (c) displays the recognition results for printed attack faces. It is evident that the DBC_ResNet18 model performs well in the real-time recognition of both real and non-real faces.

Conclusion

This study addresses the key deficiencies of existing cross-domain face anti-deception models, such as insufficient deep feature representation and weak generalization ability, by integrating multiple advanced technologies including ResNet-18 as the backbone, dynamic convolution, bottleneck concern module, central difference convolution, and adaptive instance normalization. The proposed DBC_ResNet18 model not only achieves state-of-the-art performance on four benchmark datasets (OULU-NPU, MSU-MFSD, CASIA-FASD, Rereplay Attack), It features a significantly low HTER (0.06%–6.22%) and a high AUC(97.55%–99.88%), while maintaining an efficiency of 0.0388 seconds for single-image prediction and avoiding overly complex models. In terms of the evaluation of work results, by separating content and style features and combining contrastive learning and adversarial training, the model's ability to capture activity-related details (for example, eyes, noses, and mouths visualized through Grad-CAM) has been significantly enhanced, and domain differences have been reduced. This means that integrating multimodal feature extraction and adaptive normalization techniques is a feasible direction to enhance the anti-spoofing robustness of faces across scenarios. Therefore, the DBC_ResNet18 model can be deployed in real-world scenarios that require high-security facial recognition, such as public security (for example, access control in government buildings), financial services (for example, mobile payment verification), and smart devices (for example, smartphone unlocking systems). However, when dealing with unevenly distributed data in specific domains, its performance may become unstable because feature alignment is difficult in such cases, and the model's reliance on depth maps (generating real faces through PRNet and setting false faces to zero) may limit its adaptability to scenarios where depth information is hard to obtain or there is noise. Therefore, The next work plan includes addressing the issue of unbalanced data distribution by improving samply-level feature alignment through the introduction of asymmetric instance adaptive whitening[41], expanding the compatibility of the model with more diverse types of spoofing attacks (for example, 3D mask attacks that are not fully covered in the current dataset), and exploring lightweight versions of the model So as to be able to be deployed on resource-constrained edge devices (for example, low-power surveillance cameras).

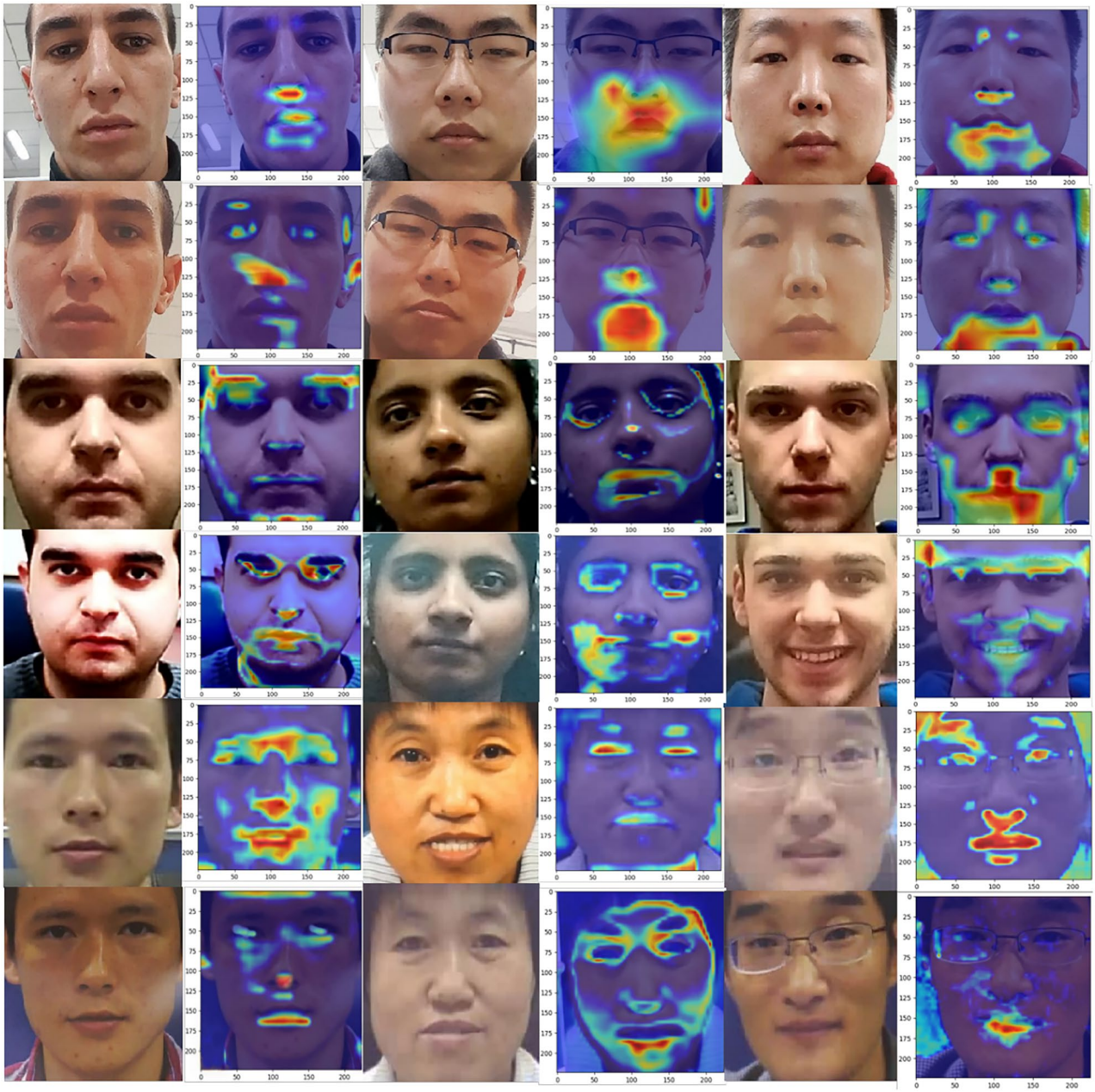
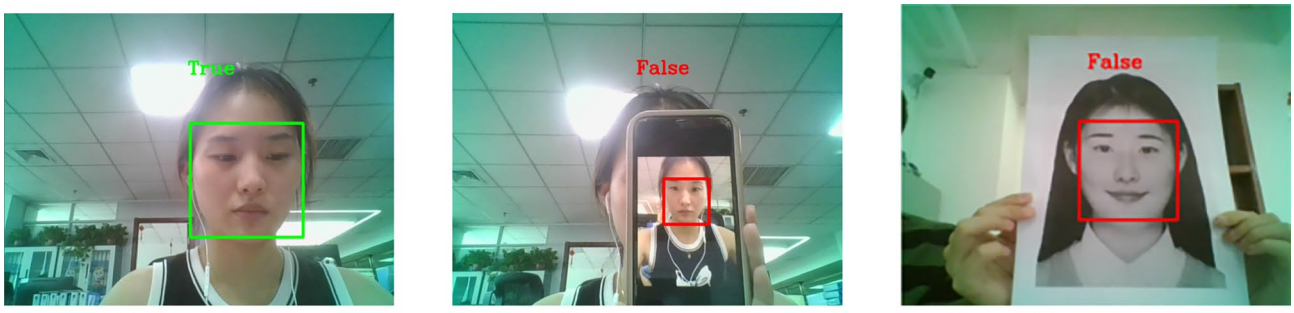


Fig. 13. The visualization of the feature maps output by the algorithm in this paper.



(a) Real human face **(b)** Mobile replay attack **(c)** Print attack

Fig. 14. Real-time face testing results.

Data availability

The dataset used in this study is available from the corresponding author upon reasonable request.

Received: 14 December 2024; Accepted: 22 September 2025
Published online: 28 October 2025

References

1. Xie, X. H., Bian, J. T., & Lai, J. H. Review of face anti-spoofing detection. *J. Image Graph.* **27**, 63–87. https://doi.org/10.11834/jig.210470 (2022).
2. Antil, A. & Dhiman, C. Unmasking deception: A comprehensive survey on the evolution of face anti-spoofing methods. *Neurocomputing* **617**, 128992. https://doi.org/10.1016/j.neucom.2024.128992 (2025).
3. Määttä, J., Hadid, A. & Pietikäinen, M. Face spoofing detection from single images using micro-texture analysis. In *2011 International Joint Conference on Biometrics*. 1–7. https://doi.org/10.1109/IJCB.2011.6117510 (2011).
4. Boulkenafet, Z., J, K. & Hadid, A. Face anti-spoofing based on color texture analysis. In *2015 IEEE International Conference on Image Processing*. 2636–2640. https://doi.org/10.1109/ICIP.2015.7351280 (2015).
5. LI, X., Komulainen, J. & Zhao, G. Generalized face anti-spoofing by detecting pulse from face videos. *2016 23rd International Conference on Pattern Recognition*. 4244–4249. https://doi.org/10.1109/ICPR.2016.7900300 (2016).
6. Yang, J., Lei, Z. & Li, S. Learn convolutional neural network for face anti-spoofing. *arXiv preprint* arXiv:1408.5601 (2014).
7. George, A. & Marcel, S. On the effectiveness of vision transformers for zero-shot face anti-spoofing. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*. 1–8. https://doi.org/10.1109/IJCB52358.2021.9484333 (2021).
8. Yu, Z., Li, X., Wang, P. & Zhao, G. Transrppg: Remote photoplethysmography transformer for 3D mask face presentation attack detection. In *2021 IEEE International Joint Conference on Biometrics (IJCB)* **28**, 1290–1294. https://doi.org/10.1109/LSP.2021.3089908 (2021).
9. Qiao, T., Wu, J., Zheng, N., Xu, M. & Luo, X. Fgdnet: Fine-grained detection network towards face anti-spoofing. *IEEE Trans. Multimed.* **25**, 7350–7363. https://doi.org/10.1109/TMM.2022.3221532 (2023).
10. Shao, R., Lan, X., Li, J. & Yuen, P. C. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10015–10023. https://doi.org/10.1109/CVPR.2019.01026 (2019).
11. Daniel, P., David, J., Artur, C. & S., R. J. L. Deep anomaly detection for generalized face anti-spoofing. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 1591–1600. https://api.semanticscholar.org/CorpusID:118717396 (2019).
12. Jourabloo, A., Liu, Y. & Liu, X. Face de-spoofing: Anti-spoofing via noise modeling. *arXiv preprint* arXiv:1807.09968 (2018).
13. Jia, Y., Zhang, J., Shan, S. & Chen, X. Single-side domain generalization for face anti-spoofing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8481–8490. https://doi.org/10.1109/CVPR42600.2020.00851 (2020).
14. Wang, Z. et al. Domain generalization via shuffled style assembly for face anti-spoofing. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4113–4123. https://api.semanticscholar.org/CorpusID:247362876 (2022).
15. Lin, J. D. et al. Defaek: Domain effective fast adaptive network for face anti-spoofing. *Neural Netw.* **161**, 83–92. https://api.semanticscholar.org/CorpusID:256292370 (2023).
16. Amin, S. U., Jung, Y., Fayaz, M., Kim, B. & Seo, S. An efficient attention-based strategy for anomaly detection in surveillance video. In *Computer Systems Science and Engineering* **46**, 3939–3958. https://doi.org/10.32604/csse.2023.034805 (2023).
17. Amin, S. U., Jung, Y., Fayaz, M., Kim, B. & Seo, S. Video anomaly detection utilizing efficient spatiotemporal feature fusion with 3d convolutions and long short-term memory modules. *Adv. Intell. Syst.* **6**, 2300706. https://doi.org/10.1002/aisy.202300706 (2024).
18. Amin, S. U., Jung, Y., Fayaz, M., Kim, B. & Seo, S. Enhanced anomaly detection in pandemic surveillance videos: An attention approach with efficientnet-b0 and cbam integration. *IEEE Access* **12**, 162697–162712. https://doi.org/10.1109/ACCESS.2024.3488797 (2024).
19. Amin, S. U., Jung, Y., Fayaz, M., Kim, B. & Seo, S. Enhancing pine wilt disease detection with synthetic data and external attention-based transformers. *Eng. Appl. Artif. Intell.* **159**, 111655. https://doi.org/10.1016/j.engappai.2025.111655 (2025).
20. Silva, V. L., Lérida, J. L., Sarret, M., Valls, M. & Giné, F. Residual spatiotemporal convolutional networks for face anti-spoofing. *J. Vis. Commun. Image Represent.* **91**, 103744. https://doi.org/10.1016/j.jvcir.2022.103744 (2023).
21. Antil, A. & Dhiman, C. Mf2shrt: Multimodal feature fusion using shared layered transformer for face anti-spoofing. *Assoc. Comput. Mach.* **20**, 6. https://doi.org/10.1145/3640817 (2024).
22. Atoum, Y., Liu, Y., Jourabloo, A. & Liu, X. Face anti-spoofing using patch and depth-based cnns. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*. 319–328. https://doi.org/10.1109/BTAS.2017.8272713 (2017).
23. Chen, Y. et al. Dynamic convolution: Attention over convolution kernels. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 11027–11036. https://doi.org/10.1109/CVPR42600.2020.01104 (2020).
24. Park, J., Woo, S. & Lee, J. Y. Learn convolutional neural network for face anti-spoofing. *arXiv preprint* arXiv:1807.06514 (2018).
25. Huang, X. & Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 1510–1519. https://doi.org/10.1109/ICCV.2017.167 (2017).
26. Yu, Z. *et al.* Searching central difference convolutional networks for face anti-spoofing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5294–5304. https://doi.org/10.1109/CVPR42600.2020.00534 (2020).
27. Yu, Z. et al. Nas-fas: Static-dynamic central difference network search for face anti-spoofing. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 43. 3005–3023. https://api.semanticscholar.org/CorpusID:226246090 (2020).
28. Ganin, Y. & Lempitsky, V. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:6755881 (2014).
29. Zhang, K., Zhang, Z., Li, Z. & Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **23**, 1499–1503. https://doi.org/10.1109/LSP.2016.2603342 (2016).
30. Feng, Y., Wu, F., Shao, X., Wang, Y. & Zhou, X. Joint 3d face reconstruction and dense alignment with position map regression network. In *European Conference on Computer Vision*. https://api.semanticscholar.org/CorpusID:3996281 (2018).
31. Boulkenafet, Z., Komulainen, J., Li, L., Feng, X. & Hadid, A. Oulu-npu: A mobile face presentation attack database with real-world variations. In *2017 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017)*. 612–618. https://doi.org/10.1109/FG.2017.77 (2017).
32. Wen, D., Han, H. & Jain, A. K. Face spoof detection with image distortion analysis. *IEEE Trans. Inf. For. Secur.* **10**, 746–761. https://doi.org/10.1109/TIFS.2015.2400395 (2015).
33. Zhang, Z. *et al.* A face antispoofing database with diverse attacks. In *2012 5th IAPR International Conference on Biometrics (ICB)*. 26–31. https://doi.org/10.1109/ICB.2012.6199754 (2012).
34. Chingovska, I., Anjos, A. & Marcel, S. On the effectiveness of local binary patterns in face anti-spoofing. In *2012 BIOSIG - Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG)*. 1–7 (2012).
35. Shao, R., Lan, X. & Yuen, P. C. Regularized fine-grained meta face anti-spoofing. In *AAAI Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:208267535 (2019).

36. Chen, Z. et al. Generalizable representation learning for mixture domain face anti-spoofing. *arXiv* **abs/2105.02453**. https://api.semanticscholar.org/CorpusID:233864896 (2021).
37. Liu, S. et al. Dual reweighting domain generalization for face presentation attack detection. *arXiv* **abs/2106.16128**. https://api.semanticscholar.org/CorpusID:235683096 (2021).
38. Liu, S. et al. Adaptive normalized representation learning for generalizable face anti-spoofing. In *Proceedings of the 29th ACM International Conference on Multimedia* (2021).
39. Zhou, Q. et al. Adaptive mixture of experts learning for generalizable face anti-spoofing. In *Proceedings of the 30th ACM International Conference on Multimedia*. https://api.semanticscholar.org/CorpusID:250699342 (2022).
40. Du, Z., Li, J., Zuo, L., Zhu, L. & Lu, K. Energy-based domain generalization for face anti-spoofing. In *Proceedings of the 30th ACM International Conference on Multimedia* (2022).
41. Zhou, Q. et al. Instance-aware domain generalization for face anti-spoofing. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20453–20463. https://api.semanticscholar.org/CorpusID:258079229 (2023).
42. He, D., Guo, H. & Li, Z. D. Face anti-spoofing detection method based on content style enhancement and feature embedding optimization. *Appl. Res. Comput.* 1869–1875. https://doi.org/10.19734/j.issn.1001-3695.2023.09.0443 (2024).
43. Wang, G., Han, H., Shan, S. & Chen, X. Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6677–6686. https://api.semanticscholar.org/CorpusID:214802592 (2020).
44. Selvaraju, R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. https://doi.org/10.1109/ICCV.2017.74 (2017).

Acknowledgements

This work has been supported by the funding provided by the National Natural Science Foundation of China (No. 61503229), the Shanxi Patent Transformation Special Programs (No. 202302009, No. 202302012), the Basic Research Program (Free Exploration) of Shanxi Province (No. 20210302123334), the Achievement Transformation and Technology Transfer Base of Taiyuan Normal University (No. 2023P003), and the Ministry of Education Humanities and Social Sciences Research Planning Fund (No. 23YJAZH118), as well as the Shanxi Province Applied Basic Research Project (No. 202103021224289).

Author contributions

The authors contributed equally to this work. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Ethical statement

(1) For Approval:this study used publicly available face liveliness detection datasets, which are free and accessible. The use of these datasets does not involve any patient privacy issues. Additionally, the real-time detection of whether the face belongs to the individual does not raise privacy concerns. We hereby declare that we have complied with the terms and conditions set by the dataset creators, ensuring that the participants' rights and privacy are not violated when handling public datasets. (2) For accordance, all methods were carried out in accordance with relevant guidelines and regulations. (3) For consent, informed consent was obtained from all subjects. (4) For consent for publication:that informed consent has been obtained from all participants for the publication of identifiable information/images in.

Additional information

Correspondence and requests for materials should be addressed to C.G.