Applied Intelligence (2025) 55:866
https://doi.org/10.1007/s10489-025-06761-2

Joint spatial-frequency deepfake detection network based on dual-domain attention-enhanced deformable convolution

Lan Qiusong[1] · Yang Chengfu[1,2] · A Qinhua[1] · Zhao Jianlong[1] · Li Jin[1]

Accepted: 28 June 2025

Abstract

This paper proposes a novel Spatio-Frequency Deepfake Detection Network based on Dual-Domain Attention and Deformable Convolution (HFDCDNet). The core of the network is the HFDCD module, which simultaneously extracts spatial domain features and high-frequency information from the input feature maps. These two types of features are then fused using a specially designed bidirectional cross-attention mechanism. Specifically, the spatial features are extracted using a Deformable Convolution and Dual Attention-based module (DCD), which leverages deformable convolutions (DCN) and a spatial-channel attention mechanism to capture more comprehensive spatial representations. Meanwhile, high-frequency features are extracted using the High-Frequency Extraction (HFE) module, which learns frequency domain representations through high-pass filtering and frequency-aware convolutional learning. The bidirectional cross-attention mechanism facilitates complementary fusion and mutual enhancement of spatial and frequency features, enabling more fine-grained and holistic feature learning and improving detection performance. To evaluate the effectiveness of the proposed HFDCDNet, extensive experiments were conducted on two widely used public datasets: FaceForensics++ and Celeb-DF (V2). The results demonstrate that HFDCDNet achieves an accuracy (ACC) of 98.31% and an area under the curve (AUC) of 99.51% on the FaceForensics++ dataset, and 98.29% ACC and 99.13% AUC on the Celeb-DF (V2) dataset, outperforming many state-of-the-art methods. These results confirm that the proposed DCD module significantly enhances the model's ability to detect manipulated facial content.

Keywords Deepfake · DeformableConv · Attention · Frequency

1 Introduction

With the rapid advancement of deep learning technologies, Deepfake techniques have found increasingly widespread applications in image and video domains. By leveraging generative adversarial networks (GANs) [1–5], Deepfake technology can realistically simulate facial expressions, lip movements, and speech, achieving highly deceptive results. These techniques have been widely applied in fields such as film special effects, entertainment, and virtual reality [6]. However, the misuse of Deepfake technology has raised a series of social concerns, including privacy violations, the spread of disinformation, and reputational damage. This has made the detection of Deepfake content an urgent task. Therefore, how to accurately and efficiently detect such forged content has become a crucial research topic in computer vision and multimedia analysis.

Although progress has been made in Deepfake detection, numerous challenges remain in practical applications.

A Qinhua, Zhao Jianlong, and Li Jin contributed equally to this work.

✉ Yang Chengfu
yangchengfu@ynnu.edu.cn

Lan Qiusong
704720234@qq.com

A Qinhua
aqinhua@ynnu.edu.cn

Zhao Jianlong
1362784075@qq.com

Li Jin
2649120532@qq.com

[1] School of Information, Yunnan Normal University, Juxian Street, Kunming 650500, Yunnan, China

[2] Engineering Research Center of Computer Vision and Intelligent Control Technology, Yunnan Provincial Department of Education, Juxian Street, Kunming 650500, Yunnan, China

Published online: 12 July 2025

First, traditional networks have limitations in feature extraction. Deepfake-generated facial images often exhibit subtle variations in local features such as texture and color, along with complex geometric deformations. Conventional convolutional neural networks (CNNs) and visual transformers (ViTs) [7] struggle to effectively capture such non-rigid transformations. Moreover, many Deepfake videos manipulate only specific facial regions, and the forged areas often display distinct characteristics across spatial and channel dimensions. However, existing CNN models lack dedicated attention mechanisms to focus on these critical regions, thereby reducing detection accuracy. Additionally, various Deepfake methods–such as Deepfakes, FaceSwap, Face2Face [8], and NeuralTextures [9]–present different types of forgery features, some of which are not apparent in the spatial domain. Consequently, detection methods relying solely on spatial-domain features (e.g., Xception [10], CapsuleNet [11]) may fail to accurately distinguish authentic from forged content.

To address these issues, we propose a novel spatial-domain feature extraction module named DCD, which integrates deformable convolutional networks (DCNs) [12] and a spatial-channel attention mechanism [13]. By introducing deformable convolutions, our method adaptively adjusts the sampling positions of convolutional kernels, enabling the network to flexibly capture traces of facial manipulation. Meanwhile, the attention mechanism helps the model focus more on facial forgery regions, effectively suppressing background noise. The proposed DCD module parallelizes dual-domain attention with deformable convolutions and fuses the resulting features, thereby optimizing spatial, channel, and geometric information simultaneously. In addition, a channel shuffling operation is applied to the fused features to redistribute inter-channel information. This disrupts the fixed associations produced by convolutional layers, enhancing representational capacity and mitigating overfitting.

Furthermore, considering that some forgery traces may be imperceptible in the spatial domain, we incorporate high-frequency information analysis. A high-frequency feature extraction module (HFE) is introduced to perform frequency-domain learning, capturing subtle artifacts embedded in the frequency spectrum. To fully leverage both spatial and frequency-domain features for Deepfake detection, we propose a joint spatio-frequency feature module, named HFDCD. In this module, input features are simultaneously processed by the DCD and HFE modules to obtain spatial and frequency representations. These are then fused using a bidirectional cross-attention mechanism, enhancing useful signals from both domains. The resulting fused features serve as the output of the HFDCD module.

2 Related work

2.1 Generation techniques

In recent years, Deepfake generation techniques have advanced significantly, primarily driven by the rapid development of deep learning, especially Generative Adversarial Networks (GANs). Currently, mainstream Deepfake generation methods include Deepfakes, FaceSwap, Face2Face, FaceShifter, and NeuralTextures [14]. These approaches utilize various techniques to manipulate facial attributes in videos, aiming to produce forged content that is often indistinguishable from authentic footage to the human eye.

Specifically, the Deepfakes method is based on autoencoders, which consist of an encoder and a decoder. The encoder transforms input facial images into low-dimensional latent representations, which the decoder then reconstructs into facial images, achieving identity swapping. The Face2Face method employs real-time facial expression capture and retargeting, transferring the facial expressions from a source actor to a target actor to generate realistic facial reenactment. FaceSwap, on the other hand, uses 3D modeling techniques to construct 3D facial models of both the source and target individuals, enabling accurate facial replacement by mapping the source face onto the target facial structure.

FaceShifter is a more advanced face-swapping approach that utilizes a dual-stage GAN-based generator to achieve high-quality identity replacement. Its two-stage architecture improves semantic consistency and visual fidelity. In contrast, NeuralTextures relies on high-resolution neural texture modeling and neural rendering. It dynamically generates photorealistic facial textures based on the subject's expressions, providing a powerful mechanism for generating convincing fake images and videos through fine-grained texture control.

2.2 Detection techniques

As Deepfake generation techniques continue to evolve, the development of effective detection methods has become a critical research focus. Modern detection approaches predominantly rely on deep learning, especially convolutional neural networks (CNNs) and vision transformers (ViTs), due to their powerful automatic feature extraction capabilities. These models aim to identify subtle artifacts or inconsistencies introduced during forgery.

For instance, in 2018, Afchar et al. proposed a lightweight model called MesoNet for Deepfake detection, achieving promising results with reduced computational complexity [15]. In 2020, Yuyang Qian et al. introduced F3-Net, which incorporates frequency domain analysis to capture high-frequency artifacts commonly lost in Deepfake

videos, significantly improving detection performance [16]. In 2021, Yuval Nirkin et al. proposed a model focusing on discrepancies between facial regions and surrounding areas, using these regional differences as auxiliary cues to enhance detection accuracy [17]. That same year, Hongshuo Chen et al. proposed the DefakeHop network, which prioritizes model efficiency and performs well with limited computational resources [18]. Shen Chen et al. introduced a forgery detection method based on local relational learning, which models interactions among facial sub-regions to improve robustness [19]. In 2022, Junyi Cao et al. presented the RECCE network, which exploits reconstruction residuals between real and fake faces to guide the learning of forgery traces [20]. Simultaneously, Zhiqing Guo et al. designed an attention mechanism specifically for manipulation traces. Their model emphasizes the most informative features via tensor refinement and attention modules, leading to high detection accuracy [21].

In 2023, Hao Lin et al. proposed a Deepfake detection model that integrates multi-scale convolutional layers and ViT. By combining dilated and depthwise separable convolutions, the model captures tampering cues across multiple spatial resolutions and leverages ViT to model global dependencies [22]. Gaoming Yang et al. observed that manipulated videos exhibit optical flow discontinuities and proposed FDS_2D, a model that exploits this inconsistency for detection [23]. Sayantan Das et al. developed a spatio-temporal detection framework that integrates both spatial and temporal cues. The model employs a dual-branch architecture and incorporates a masking mechanism during training to enhance attention on manipulated regions [24].

In 2024, Ankit Yadav et al. pointed out that equal-weighted fusion of multi-scale features may lead to sub-optimal performance. They introduced Face-NeSt, a novel architecture that adaptively selects fusion ratios for multi-scale features, thereby enhancing detection effectiveness [25].

Despite these significant advances, current Deepfake detection models still face several challenges, including limited sensitivity to subtle geometric deformations, susceptibility to interference from non-manipulated regions, and insufficient integration of frequency domain information. Therefore, designing a robust detection approach that combines spatial-domain artifact localization with frequency-domain analysis is of great importance. Such hybrid models can significantly enhance detection accuracy and reliability, making them highly valuable in applications such as surveillance, digital forensics, and media authentication.

3 Methodology

3.1 DCD module

To enhance the network's ability to extract spatial features for Deepfake detection, we propose a novel spatial-domain feature extraction module named the DCD module, which integrates Deformable Convolutional Networks (DCN) [12], spatial-channel attention mechanisms [13], and feature fusion strategies. In this section, we first introduce the key components–DCN and the spatial-channel attention mechanism,and then describe the architecture of the proposed DCD module in detail.

3.1.1 Deformable convolution

Deformable convolution is an advanced convolutional architecture that addresses the limitations of traditional convolution operations. Unlike standard convolution, which applies fixed geometric sampling locations, deformable convolution introduces learnable offsets that allow the convolution kernel to adapt dynamically to the shapes and structures present in the input feature map [12]. This flexibility significantly enhances the ability of the network to model spatial deformations and irregularities in images.

In the context of Deepfake detection, deformable convolution exhibits strong application potential. Facial regions often exhibit complex structures and irregular spatial arrangements in images, especially in forged samples. Standard convolution kernels, constrained by their fixed receptive fields, struggle to effectively capture such spatial inconsistencies or subtle manipulations. In contrast, deformable convolution can adjust its sampling locations based on the input features, thereby enabling more precise extraction of local facial details, such as around the eyes, mouth corners, or facial contours–areas that are frequently tampered with in Deepfake content.

By incorporating deformable convolution into the detection framework, the model becomes more adaptable and robust, especially in handling non-rigid distortions and spatial artifacts typically introduced in forged face images. A conceptual illustration of deformable convolution is provided in Fig. 1.

As shown in Fig. 1, the core concept of deformable convolution lies in transforming the fixed and rigid sampling grid of a standard convolutional kernel into a more flexible one that can dynamically adjust according to the input features. In its implementation, deformable convolution retains the basic computational operations of standard convolution, but incorporates a learnable offset parameter that modifies the sampling positions within the receptive field. This modification enables the kernel to adapt to irregular spatial

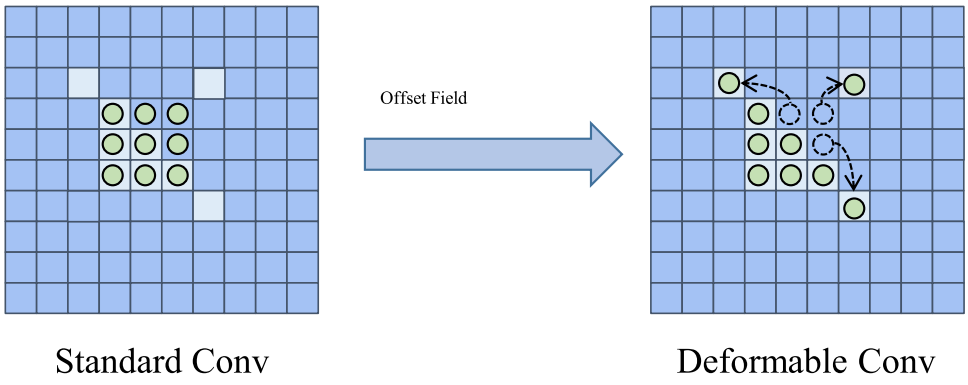


Fig. 1 Schematic diagram of deformable convolution

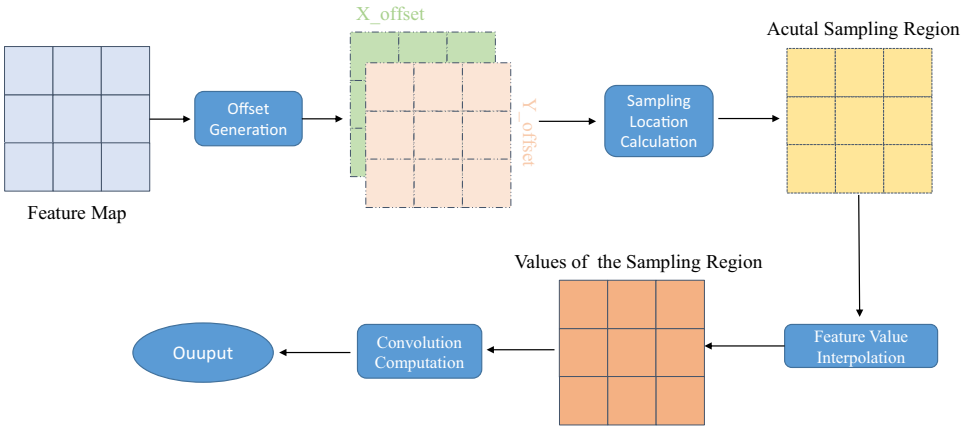


Fig. 2 Deformable convolution calculation process

patterns while maintaining the general structure of conventional convolution. The overall workflow of deformable convolution is illustrated in Fig. 2.

As depicted in Fig. 2, the actual implementation of deformable convolution involves the following steps: First, offset fields are learned to determine the sampling displacements in both the X and Y directions for each location in the convolution window. These offsets are then used to compute the precise sampling coordinates. Next, the feature values at these offset positions are obtained, typically through bilinear interpolation. Finally, these sampled values are used in the convolution operation to produce the output feature map.

3.1.2 Spatial-channel attention

In Deepfake detection, models are often susceptible to interference from non-manipulated regions, which can degrade detection performance. An effective attention mechanism can mitigate this issue by enhancing the model's focus on critical regions–areas that typically correspond to manipulated parts of the image. To this end, we incorporate an attention module that combines both spatial and channel dimensions to guide the model towards regions that are more likely to contain tampering artifacts [13].

This attention mechanism consists of two main components: the Channel Attention Module (CAM) and the Spatial Attention Module (SAM). These components perform attention operations along the channel and spatial dimensions, respectively, allowing the model to emphasize informative spatial features while dynamically adjusting the channel-wise responses of the feature maps. The overall structure of this attention mechanism is illustrated in Fig. 3. As shown in Fig. 3, the input feature map is element-wise multiplied by two sets of attention weights: one for channel attention and the other for spatial attention. Specifically, to compute the channel attention, both average pooling and max pooling are applied independently to the input feature map, and the resulting feature descriptors are passed through a shared Multi-Layer Perceptron (MLP). The outputs are then summed and activated to generate the channel attention weights. For spatial attention, the average-pooled and max-pooled features are concatenated along the channel axis and passed through a convolution layer followed by a non-linear activation function to produce the spatial attention weights.

This dual attention design enables the model to adaptively focus on semantically meaningful regions and channels, thereby enhancing the robustness and accuracy of Deepfake detection.

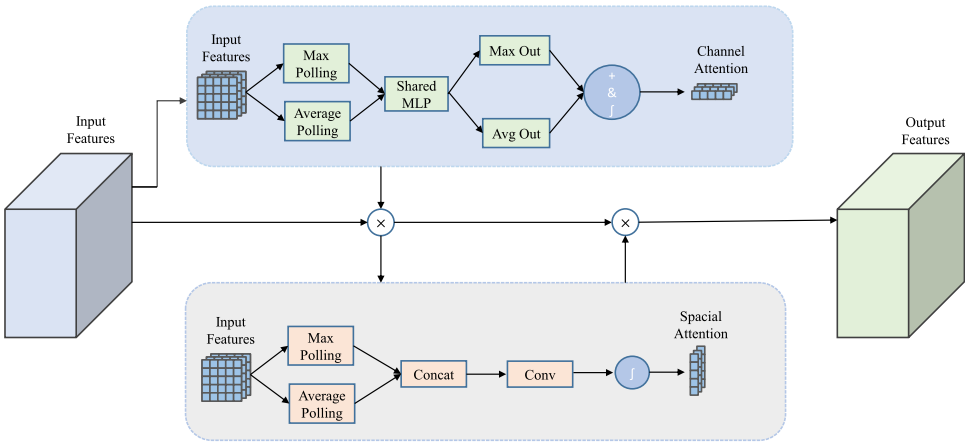


Fig. 3 Spatial-Channel Attention Architecture

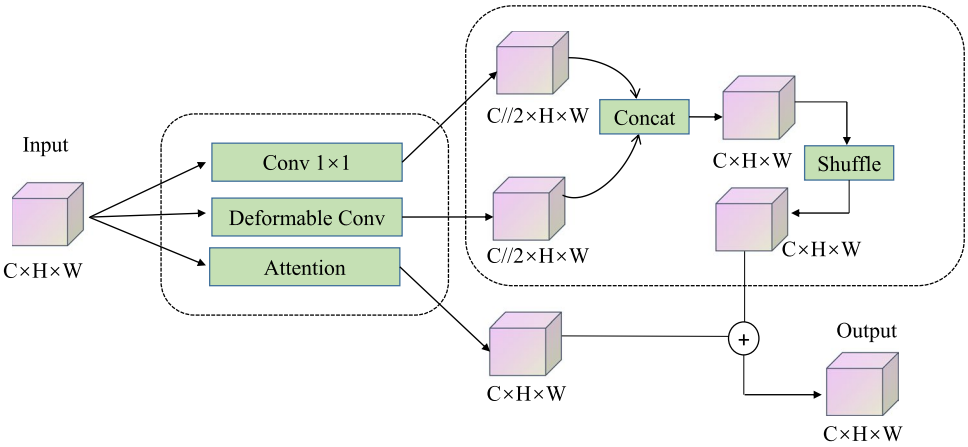


Fig. 4 DCD Module

3.1.3 DCD module architecture

To flexibly integrate the aforementioned attention mechanism and deformable convolution for deepfake detection, we propose a novel module termed the DCD module. Rather than simply stacking deformable convolution and attention modules, the DCD module employs a more sophisticated design: the input feature map is simultaneously fed into a standard convolution branch, a deformable convolution branch, and an attention module. The outputs from these branches are then fused and shuffled. This approach enables the network to capture both key regions and non-rigid deformations occurring during manipulation at the same scale, thereby enriching the feature representation capability and effectively mitigating overfitting. The overall workflow of the DCD module is illustrated in Fig. 4.

Specifically, let the input feature map to the DCD module be denoted as X, with dimensions (B,C1,H,W), where B represents the batch size, C1 represents the number of channels, and H and W represent the height and width of the feature map, respectively. The output feature map of the DCD module is denoted as Y, with the same dimensions as X, i.e., (B,C1,H,W). To derive Y from X, a series of operations are applied, including standard convolution, deformable convolution, feature fusion and shuffling, and two-dimensional attention computation. Initially, X is input in parallel to a standard convolution module, a deformable convolution module, and an attention mechanism. This can be represented as:

$$X_{\text{conv}} = \text{Conv}(X) \tag{1}$$

$$X_{\text{Dconv}} = \text{DConv}(X) \tag{2}$$

$$Y_{\text{attn}} = \text{Attention}(X) \tag{3}$$

where Conv represents the standard convolution operation, DConv denotes deformable convolution, and Attn refers to the two-dimensional attention mechanism. The resulting feature maps X_{conv}, X_{Dconv}, and Y_{attn} have dimensions of (B,C1//2,H,W), (B,C1//2,H,W) and (B,C1,H,W), respectively.

Next, the feature maps X_{conv} and X_{Dconv} are concatenated along the channel dimension and subjected to a channel shuffling operation, generating Y_{conv}. This process facilitates cross-channel information exchange and integration within the spatial dimension:

$$Y_{\text{conv}} = \text{Shuffle}\left(\text{Concat}(X_{\text{conv}}, X_{\text{Dconv}})\right) \tag{4}$$

where Concat denotes concatenation along the channel dimension, and Shuffle represents the channel shuffling operation. The resulting Y_{conv} has dimensions (B,C1,H,W).

Finally, the outputs of the attention mechanism, Y_{attn}, and the output from the channel concatenation and shuffling, Y_{conv}, are fused to obtain the final output Y. Since both Y_{attn} and Y_{conv} have the same dimensions, i.e., (B,C1,H,W), they can be directly added to achieve feature fusion:

$$Y = Y_{\text{attn}} + Y_{\text{conv}} \tag{5}$$

Thus, the final output Y of the DCD module incorporates information extracted from the input feature map X through standard convolution, deformable convolution, and the two-dimensional attention mechanism. Compared to traditional methods, this approach provides a more comprehensive and richer representation of spatial domain features.

3.2 HFDCD module

As mentioned earlier, some forgery methods leave subtle or indistinct traces in the spatial domain, making it challenging to detect them by focusing solely on spatial information. Therefore, based on the DCD module, we further propose the HFDCD module, which simultaneously emphasizes spatial and high-frequency information within the feature maps, thereby providing a more comprehensive and effective feature representation. In the HFDCD module, spatial features are extracted using the DCD module, while frequency-domain features are obtained through a High-Frequency Feature Extraction (HFE) module. These two types of features are then fused using a specially designed bidirectional cross-attention mechanism. In the following, we introduce the HFE module, the bidirectional cross-attention module, and the overall architecture of the HFDCD module.

3.2.1 HFE module

The HFE module is responsible for extracting frequency-domain information from the input feature maps. It mainly consists of two components: high-frequency processing and frequency-domain learning. The high-frequency processing step forces the model to focus exclusively on high-frequency information within the frequency feature space, thereby reducing interference from low-frequency components and preventing overfitting to specific frequency patterns.

Applying high-frequency processing along the channel dimension of the input features has demonstrated positive effects in deepfake detection tasks [26]. This is because each channel of a feature map essentially encodes an abstract semantic representation of the image. In natural images, semantic relationships and frequency responses across channels tend to be highly coordinated and consistent. However, forged images often exhibit frequency inconsistencies, spectral perturbations, or abrupt changes in channel information. Such abrupt changes correspond to high-frequency components in the frequency domain. Thus, preserving high-frequency information along the channel dimension helps reveal inconsistencies between channels, which is critical for effective deepfake detection.

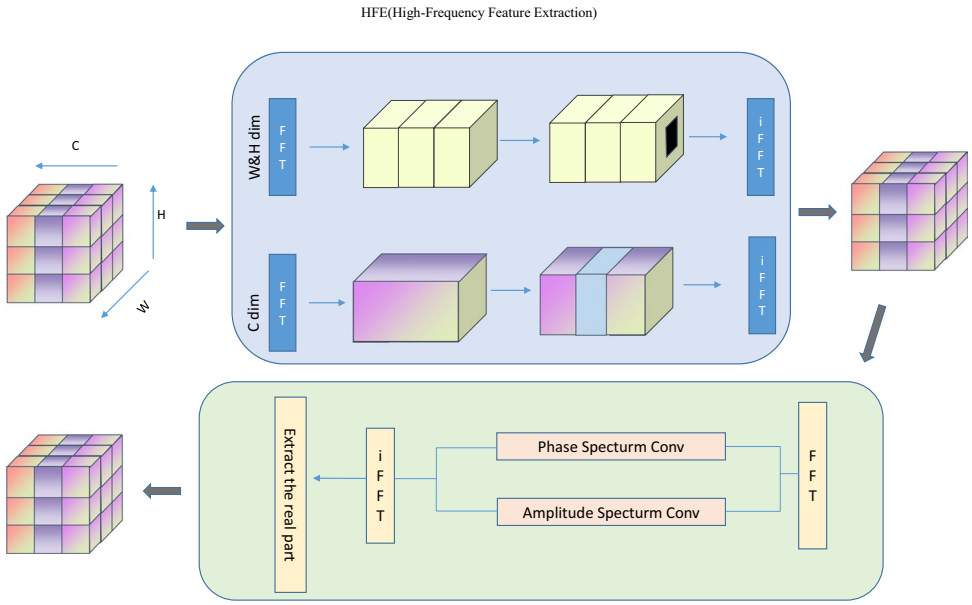
Furthermore, since the scale and semantic content of feature maps evolve with network depth, the HFE module enables multi-scale fusion of high-frequency information, thereby improving detection accuracy and efficiency. Through frequency-domain learning, the model can extract discriminative features in the frequency spectrum and further suppress irrelevant frequency-domain noise, resulting in key frequency features that better support the model's decision-making process.

The detailed workflow of the HFE module is illustrated in Fig. 5.

As illustrated in Fig. 5, the HFE module extracts frequency-domain information from the input feature map through the following steps: assuming the input feature map X has dimensions (H,W,C), a Fast Fourier Transform (FFT) is first applied along both the spatial dimensions (H,W) and the channel dimension C, converting the feature map from the spatial domain to the frequency domain. The zero-frequency component is then shifted to the center, bringing the low-frequency components closer to the center while pushing high-frequency components to the periphery. To emphasize high-frequency information, a masking operation is employed to suppress low-frequency components by zeroing them out, thereby retaining only the high-frequency signals. Subsequently, these high-frequency components are transformed back into the spatial domain via an inverse FFT (IFFT), and only the real part of the resulting output is preserved for further processing. This treatment is justified because, although both real and imaginary parts are involved in the inverse transformation, the imaginary part may still retain non-zero values after IFFT. These values often result from high-frequency preservation and do not correspond to actual phase information; instead, they may introduce noise and adversely affect model training. Moreover, since real-valued images inherently lack an imaginary component, it is both natural and necessary to ensure that the reconstructed feature maps after high-frequency filtering reside in the real-valued domain.

Following this high-frequency processing, the resultant high-frequency features undergo frequency-domain learning. Specifically, the processed features are again transformed to the frequency domain, where both the real and

Fig. 5 HFE Module



imaginary parts are separately learned via convolutional layers. The learned components are then combined and converted back to the spatial domain, and similarly, only the real part is retained as the final feature representation.

The proposed HFE module leverages FFT to extract high-frequency information from input features, combining high-frequency filtering and frequency-domain learning to focus on detecting spectral anomalies associated with forged images, such as frequency inconsistencies across channels.

Notably, the well-known frequency-based model F3-Net [16] also exploits frequency-domain information for forgery detection by employing Frequency Aware Decomposition (FAD) and Local Frequency Statistics (LFS) to capture frequency features, emphasizing amplitude spectrum differences fused with a CNN branch. Although both approaches aim to leverage frequency information, the HFE module distinguishes itself by performing high-frequency processing along the channel dimension, thereby capturing inter-channel semantic frequency inconsistencies to aid the detection network's decision-making. In contrast, F3-Net's FAD and LFS focus mainly on single-scale frequency statistics and do not explicitly model dynamic inter-channel relationships.Moreover, unlike F3-Net, which applies FFT to the entire frequency spectrum to extract features, the HFE module first suppresses low-frequency components to highlight high-frequency anomalies in forged regions and then applies frequency-domain convolutional layers for deep feature modeling. This targeted approach significantly enhances the discriminative power of frequency-domain features.

3.2.2 Bidirectional cross-attention

In recent years, multimodal fusion research has increasingly incorporated information-theoretic modeling strategies to achieve disentangled representations of shared and modality-specific features. A representative work by Thao et al. [27], named MEDFuse, addresses multimodal data modeling in electronic health records (EHR) by jointly encoding clinical text and structured laboratory test data. This method leverages large language models (LLMs) to extract textual embeddings, trains laboratory data representations via masking mechanisms, and designs a disentangled Transformer architecture combined with mutual information constraints to model both shared and distinct features across modalities. However, while such strategies excel in heterogeneous modality fusion tasks–such as integrating text and tabular data–they are not directly applicable to deepfake image detection. First, MEDFuse handles modalities with substantial semantic differences, relying heavily on context-aware semantic alignment, whereas deepfake detection requires capturing fine-grained feature variations in both spatial and frequency domains of images.

To address this need, this paper proposes a bidirectional cross-attention mechanism to explicitly fuse complementary information between spatial and frequency branches, effectively capturing subtle forgery traces introduced by manipulations in both domains.

Within this module, spatial and frequency information are cross-attended and fused through a uniquely designed bidirectional cross-attention mechanism. In the spatial-to-frequency cross-attention, spatial features guide the attention toward salient frequency components. Conversely, frequency-to-spatial cross-attention enables frequency

features to focus on critical spatial regions. This bidirectional mechanism allows mutual enhancement between spatial and frequency information, leading to more detailed and comprehensive feature learning that better captures forgery artifacts in images. Specifically, the output of the DCD module is used as the spatial domain feature, while the output of the HFE module serves as the frequency domain feature. Each feature is processed through convolutional layers to produce their respective Query (Q), Key (K), and Value (V) matrices. The spatial domain Q is then used to attend over the frequency domain K and V, while the frequency domain Q attends over the spatial domain K and V. This results in an enhanced frequency feature guided by key spatial features and an enhanced spatial feature guided by important frequency features.

Through this bidirectional cross-attention, the model collaboratively analyzes forgery artifacts from two perspectives: inferring potential forged regions from frequency anomalies and inferring critical frequency components from spatial semantic structures. Finally, these two enhanced features are fused via learnable weighting parameters, forming a mutually reinforcing detection mechanism for spatial and frequency information. To further improve the fusion's effectiveness and detection performance, two learnable weight matrices are employed to weight and combine the enhanced frequency and spatial features, respectively. The overall process is illustrated in Fig. 6.

As shown in Fig. 6, let the spatial domain features output by the DCD structure be denoted as X_s, and the features obtained through high-frequency processing and frequency domain learning as X_F. These two feature sets, X_s and X_F, will be separately processed by convolution operations to generate their corresponding query (Q), key (K), and value (V) matrices:

$$Q_s, K_s, V_s = Q_{\text{conv}}(X_s), K_{\text{conv}}(X_s), V_{\text{conv}}(X_s) \tag{6}$$

where Q_s, K_s, V_s are the query, key, and value matrices derived from the spatial domain features X_s, and Q_F, K_F, V_F are the query, key, and value matrices derived from the frequency domain features X_F. The operations Q_{conv},K_{conv} and V_{conv} represent the convolution operations used to generate the respective matrices. Next, the attention computations are performed to obtain the enhanced spatial features and enhanced frequency domain features:

$$Attention_{sf} = \text{Softmax}\left(\frac{Q_s K_f^\top}{\sqrt{d}}\right) V_f \tag{7}$$

$$Attention_{fs} = \text{Softmax}\left(\frac{Q_f K_s^\top}{\sqrt{d}}\right) V_s \tag{8}$$

where $Attention_{sf}$ and $Attention_{fs}$ represent the results of cross-attention from spatial domain to frequency domain and from frequency domain to spatial domain, respectively. These two attention results are multiplied by corresponding learnable weight matrices and subsequently fused to obtain the final feature map:

$$X_{out} = W_{sf} \cdot Attention_{sf} + W_{fs} \cdot Attention_{fs} \tag{9}$$

In the equation, X_{out} represents the final output, while W_{sf} and W_{fs} denote the weight matrices for spatial-to-frequency cross-attention and frequency-to-spatial cross-attention, respectively.

Compared to simple concatenation and unidirectional attention mechanisms, the proposed bidirectional cross-attention module offers superior capabilities in information

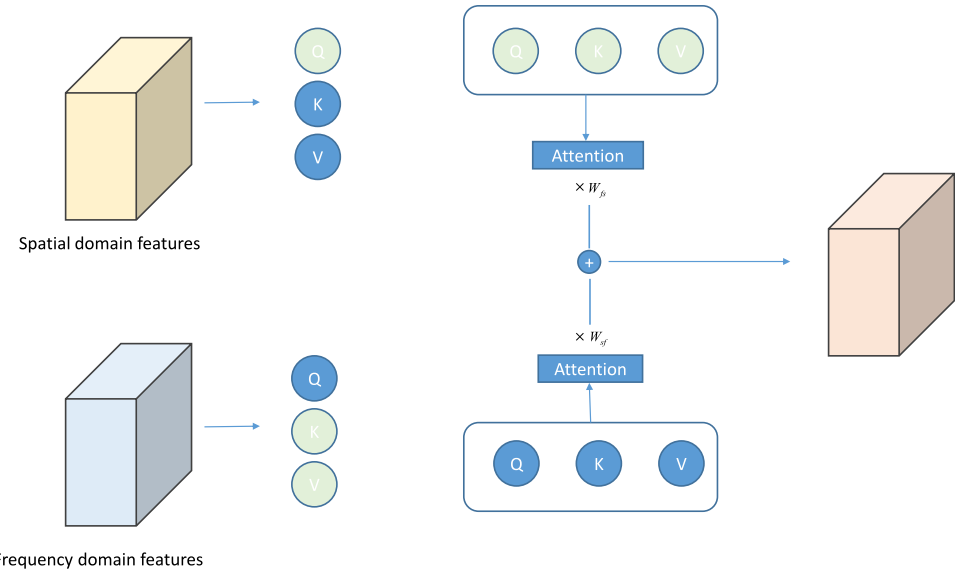


Fig. 6 Bidirectional Cross-Attention Module

complementation and interactive modeling. When using only concatenation or unidirectional attention, there is a higher risk of uneven feature dominance, biased fusion, or redundant representations. Similarly, multi-branch attention mechanisms often compute attention independently across different dimensions or perform unidirectional fusion (e.g., feature concatenation in F3-Net) to integrate multimodal information, which can also lead to feature imbalance and redundancy. In contrast, our bidirectional cross-attention explicitly facilitates interaction–allowing spatial features to guide frequency-domain features and vice versa–so that salient information is mutually enhanced before fusion. This dynamic complementarity between spatial and frequency features significantly improves the model?s discriminative capacity.

3.2.3 HFDCD module architecture

To efficiently integrate both spatial and frequency domain information for forgery detection, we propose the HFDCD module. In this module, the input feature map is concurrently fed into both the DCD and HFE modules to extract spatial domain and frequency domain features, respectively. Subsequently, the extracted spatial and frequency domain features are simultaneously input into the aforementioned bidirectional cross-attention module to achieve deep fusion of these features. The architecture of the HFDCD module is illustrated in Fig. 7.

As shown in Fig. 7, let the input feature map to the HFDCD module be denoted as X, with dimensions (B,C1,H,W) where B represents the batch size, C1 represents the number of channels, and H and W represent the height and width of the feature map, respectively. The corresponding output feature map of the HFDCD module is denoted as Y, with the same dimensions (B,C1,H,W). First, the feature map X is concurrently input into the DCD and HFE modules to obtain the learned spatial domain features and frequency domain features:

$$X_s = \mathrm{DCD}(X) \tag{10}$$

$$X_f = \mathrm{HFE}(X) \tag{11}$$

where X_s and X_f represent the learned spatial and frequency domain features, respectively, and their dimensions remain consistent with that of X, i.e., (B,C1,H,W). Here, DCD and HFE refer to the relevant computations within the DCD and HFE modules mentioned previously. Next, X_s and X_f are simultaneously passed through the aforementioned bidirectional cross-attention module for feature fusion:

$$Y = \mathrm{BCAttention}(X_s, X_f) \tag{12}$$

where Y represents the final output of the HFDCD module, and its dimensions remain (B,C1,H,W). It is worth noting that the input and output dimensions of the HFDCD module, as well as the previously introduced DCD module, are consistent. This careful design ensures that both the DCD and HFDCD modules can be easily integrated and transferred into various detection networks without the need for complex dimensional alignment operations.

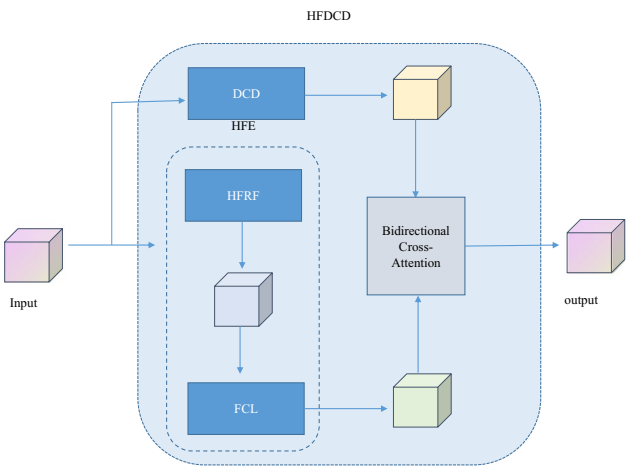


Fig. 7 HFDCDModule

3.3 HFDCD network architecture

To apply the proposed HFDCD module to the deepfake detection task, we design a novel detection network based on the HFDCD module, termed the HFDCD network. This network extends the HFDCD module by incorporating downsampling, residual connections, feature fusion, and a classifier, thus forming a complete detection architecture. The detailed structure is illustrated in Fig. 8.

As shown in Fig. 8, the proposed HFDCD network consists of three main stages: downsampling, feature extraction, and classification. Among these, the feature extraction stage primarily relies on the repeated utilization of the HFDCD modules.

4 Experiments

To verify the effectiveness of the proposed HFDCD method, we designed the HFDCD network based on this approach and conducted extensive and rigorous experiments on two publicly available and authoritative face forgery detection datasets: FaceForensics++ [14] and Celeb-DF (V2) [26]. In the following sections, we present the experimental setup, results, and detailed analyses.

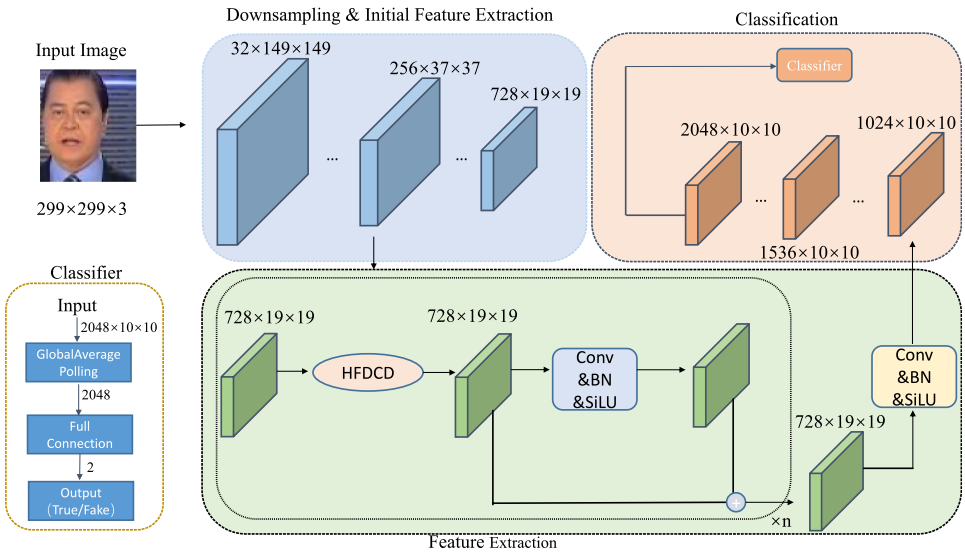


Fig. 8 HFDCDNet Architecture

Table 1 Detection indicators of HFDCD on different counterfeiting methods

Method	Accuracy	Precision	Recall	AUC
DeepFakes	99.21%	99.53%	98.89%	99.85%
Face2Face	99.01%	99.71%	98.31%	99.91%
FaceShifter	99.11%	98.26%	99.99%	99.99%
FaceSwap	98.89%	99.78%	98.00%	99.38%
NeuralTextures	97.56%	96.75%	98.43%	99.24%
Average	98.76%	98.81%	98.72%	99.67%
Celeb-DF	98.94%	99.64%	97.26%	99.58%

4.1 Datasets

4.1.1 FaceForensics++ dataset

To comprehensively evaluate the detection capability of the HFDCD network, we selected two widely used and representative datasets:

FaceForensics++ Released by Andreas et al. in 2019, FaceForensics++ has become a standard benchmark for face forgery detection. The dataset comprises 1,000 pristine videos along with 5,000 manipulated videos generated by five popular forgery methods: Deepfakes, Face2Face, FaceShifter, FaceSwap, and NeuralTextures. To facilitate investigation into the effect of video compression on detection performance, the dataset provides three compression levels: uncompressed (c0), lightly compressed (c23), and heavily compressed (c40). Our experiments adopt the widely used c23 compression version for both training and testing.

Celeb-DF (v2) This dataset contains 590 real videos and 5,639 corresponding forged videos. Compared with version 1, Celeb-DF (v2) significantly improves the quality of forged videos by addressing many flaws present in the earlier version, thereby increasing the detection challenge. Therefore, we use Celeb-DF (v2) for our experiments.

4.2 Experimental results and discussion

We trained and tested the proposed HFDCD network on the five forgery methods of the FaceForensics++ dataset as well as on the Celeb-DF (v2) dataset. The performance metrics include Accuracy, Precision, Recall, and the Area Under the Receiver Operating Characteristic Curve (AUC). The detailed results are summarized in Table 1.

As shown in Table 1, the HFDCD method achieves excellent performance across all five forgery categories on the FaceForensics++ dataset, with accuracy exceeding 97%, reaching as high as 99.21%; precision peaking at 99.78%; recall up to 99.99%; and AUC values as high as 99.99%. On average, across all five forgery types, the model attains an accuracy of 98.76%, precision of 98.79%, recall of 98.72%, and AUC of 99.67%. On the Celeb-DF (v2) dataset, the network achieves an accuracy of 98.29% and an AUC of 99.13%. These results demonstrate that the proposed HFDCD network effectively captures complementary forgery traces from both spatial and frequency domains, enabling robust detection of forged faces. Moreover, the method consistently delivers superior detection performance on two widely recognized benchmark datasets.

The confusion matrix provides a detailed overview of the model's performance across different classes, clearly illustrating its predictive capability and the misclassification patterns in the binary real-versus-fake classification task. Evaluation metrics such as accuracy, precision, and recall can all be derived from the four fundamental components of the confusion matrix: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). In the following, we comprehensively assess the proposed

deepfake detection model using confusion matrices corresponding to different forgery methods, as shown in Fig. 9.

Figure 9 presents the confusion matrices of the HFDCD network for each specific forgery method, where the x-axis represents the predicted labels and the y-axis denotes the ground-truth labels. As shown in Fig. 9, the HFDCD network accurately classifies the vast majority of images into their correct categories across all types of forgery methods. Among these, the NeuralTextures method yields the lowest performance; however, even in this case, only 0.79% of the 7,000 real images are incorrectly classified as fake, and 1.66% of the 7,000 fake images are misclassified as real–resulting in a total misclassification rate of less than 3%.

In contrast, for better-performing methods, the network misclassifies merely 0.01% of real images as fake in the case of FaceShifter, and only 0.11% of fake images as real in the case of FaceSwap. These results demonstrate the outstanding discriminative capability of the proposed HFDCD network, particularly in detecting facial forgeries with high precision and robustness.

These results indicate that the proposed HFDCD network possesses excellent discriminative capability for deepfake face detection, consistently demonstrating robust performance across diverse forgery methods.

To further validate the effectiveness of the proposed HFDCD method, we conducted comparative experiments against several state-of-the-art and representative models on the benchmark datasets FaceForensics++ and Celeb-DF (v2). The detection performance was evaluated using two key metrics: Accuracy (ACC) and the Area Under the Receiver Operating Characteristic Curve (AUC). For each forgery method, the highest ACC and AUC values achieved among all compared methods are highlighted in bold to emphasize the best performance.

The comparative results of ACC on the FaceForensics++ dataset are summarized in Table 2, while the AUC comparison results for the same dataset are presented in Table 3. The ACC and AUC results on the Celeb-DF (v2) dataset are collectively reported in Table 4.

From Tables 2 and 3, it can be observed that the proposed HFDCD network achieves strong performance across five different manipulation methods in the FaceForensics++ dataset. Notably, the HFDCD network attains the highest accuracy among all compared methods on every forgery type. Particularly on the NeuralTextures manipulation–which is challenging for most models–the HFDCD network surpasses the second-best method, AW-MSA [25], by 6 percentage points, reaching an accuracy of 97.56%. Moreover, the overall accuracy of HFDCD on FaceForensics++ reaches 98.76%. Regarding the AUC metric, the proposed HFDCD network exceeds 99% across all forgery methods, with a comprehensive AUC of 99.67%. It achieves the highest AUC for all manipulation methods except FaceSwap, where it ranks second, trailing the top-performing AW-MSA

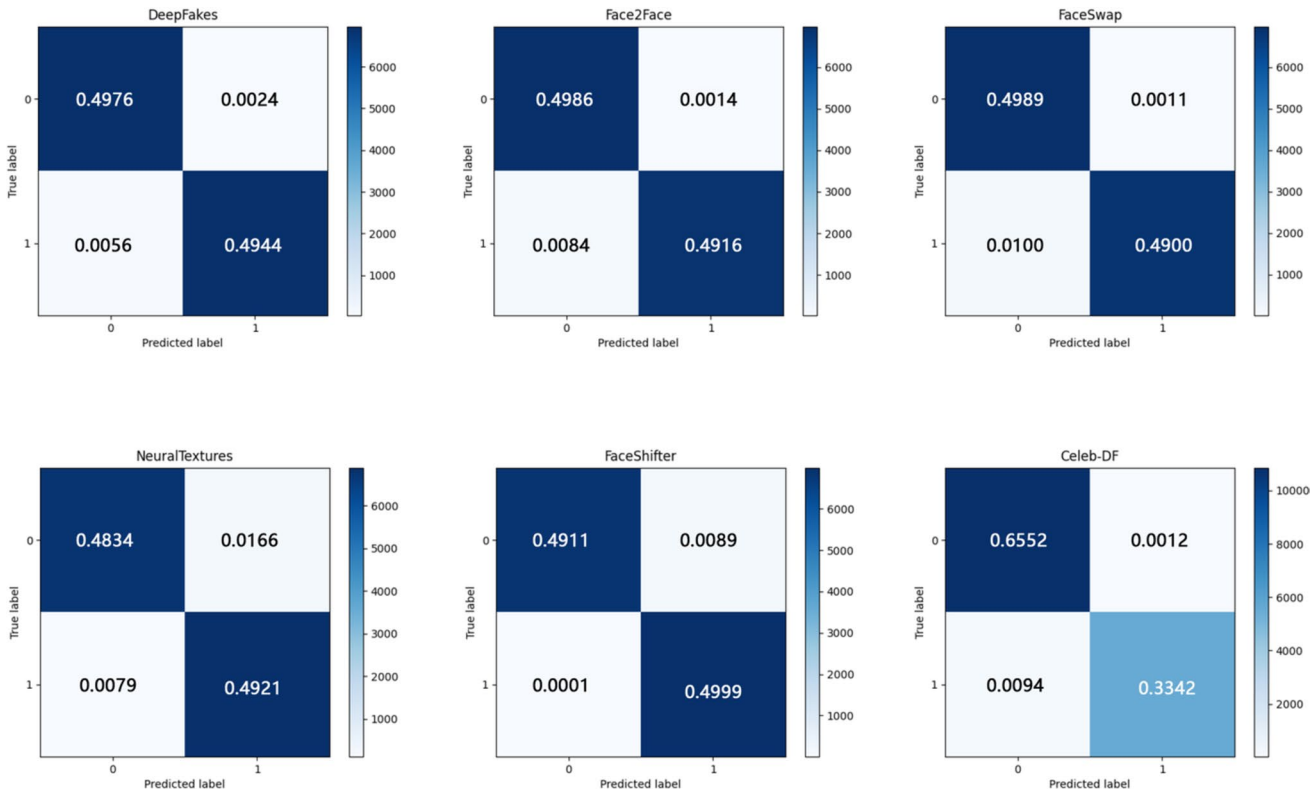


Fig. 9 Confusion matrix of HFDCDNET on different forgery methods

Table 2 Comparison of ACC indicators of FaceForensics++ dataset

Methods	Neural Texture	FaceShifter	FaceSwap	Face2Face	DeepFakes	Average
Faces&Context [17]	74.00%	-	84.50%	80.30%	94.50%	-
DeepfakeHop [18]	-	-	-	-	-	-
PRRNet [29]	80.01%	-	94.93%	90.15%	95.63%	-
FDS_2D [23]	-	-	-	-	-	-
Gradient Based [30]	-	-	-	-	-	-
TP&MTA [21]	-	-	-	-	-	-
MesoNet [15]	-	-	-	-	-	83.10%
Multi-scale Conv&ViT [22]	-	-	-	-	-	90.74%
F3-Net [16]	83.32%	-	96.53%	95.32%	95.97%	92.78%
Masked Relation Learning [31]	-	-	-	-	-	93.82%
MRT-Net [32]	90.25%	96.11%	96.76%	97.67%	96.70%	95.50%
Xception [14]	-	-	-	-	-	95.73%
AW-MSA [25]	91.28%	98.54%	97.79%	97.60%	98.05%	96.65%
E2E Learning [20]	-	-	-	-	-	97.06%
Local Relation Learning [19]	-	-	-	-	-	97.59%
MADD [33]	-	-	-	-	-	97.60%
Self.Info [34]	-	-	-	-	-	97.64%
MASDT [24]	-	-	-	-	-	98.19%
Self-Supervised [35]	-	-	-	-	-	98.31%
HFDCD (Proposed)	**97.56%**	**99.11%**	**98.89%**	**99.01%**	**99.21%**	**98.76%**

Table 3 Comparison of AUC indicators of FaceForensics++ dataset

Methods	Neural Texture	FaceShifter	FaceSwap	Face2Face	DeepFakes	Average
Faces&Context [17]	-	-	-	-	-	-
DeepfakeHop [18]	-	-	97.98	-	95.95%	-
PRRNet [29]	-	-	-	-	-	-
FDS_2D [23]	-	-	-	-	-	87.09%
Gradient Based [30]	-	-	-	-	-	90.34%
TP&MTA [21]	-	-	-	-	-	97.55%
MesoNet [15]	-	-	-	-	-	-
Multi-scale Conv&ViT [22]	-	-	-	-	-	94.86%
F3-Net [16]	-	-	-	-	-	-
Masked Relation Learning [31]	-	-	-	-	-	98.27%
MRT-Net [32]	96.62%	99.42%	99.33%	99.74%	99.10%	98.84%
Xception [14]	-	-	-	-	-	-
AW-MSA [25]	95.48%	99.78%	**99.48%**	99.05%	99.45%	98.64%
E2E Learning [20]	-	-	-	-	-	99.32%
Local Relation Learning [19]	-	-	-	-	-	99.46%
MADD [33]	-	-	-	-	-	99.29%
Self.Info [34]	-	-	-	-	-	99.35%
MASDT [24]	-	-	-	-	-	99.76%
Self-Supervised [35]	-	-	-	-	-	98.31%
HFDCD (Proposed)	**99.24%**	**99.99%**	99.38%	**99.91%**	**99.85%**	**99.67%**

[25] by only 0.1 percentage points. The overall AUC difference compared to MASDT [24] is also less than 0.1 percentage points.

As shown in Table 4, on the Celeb-DF dataset, the HFDCD network achieves an accuracy of 98.94%, the highest among all compared methods, and an AUC of 99.58%, ranking second only to Xception [14]. However, the gap between them is less than 0.2 percentage points, with both models exceeding 99.5% AUC. Notably, the HFDCD network surpasses the Xception [14] network by nearly two percentage points in accuracy. Taken together, these results demonstrate that the proposed HFDCD network achieves state-of-the-art performance across both datasets, confirming its strong detection capability for various forgery methods and verifying its effectiveness in deepfake detection tasks.

Beyond accuracy on benchmark datasets, model generalization is a critical metric in deepfake detection. Given the rapid evolution of forgery techniques, a single dataset

Table 4 Comparison of ACC and AUC indicators of Celeb-DF dataset

Methods	ACC	AUC
Faces&Context [17]	-	66.00%
TP&MTA [21]	-	67.43%
MSF [36]	-	69.50%
Defakehop [18]	-	87.65%
Gradient Based [30]	-	93.32%
DF-UDetector [37]	88.92%	-
Certainty-based Attention network [38]	92.00%	94.00%
AMTEN [39]	92.54%	88.04%
Facial Semantic Segmentation&Attention [40]	95.75%	-
F3-Net [16]	95.95%	98.93%
AW-MSA [25]	96.12%	98.23%
Xception [14]	97.90%	99.73%
MASDT [24]	98.00%	98.90%
MRT-Net [32]	98.15%	99.21%
HFDCD (Proposed)	**98.94%**	**99.58%**

Table 5 Comparison of AUC performance across different methods

Methods	AUC
Xception [14]	48.21%
CapsuleNet [11]	57.53%
M2TR [41]	63.73%
F3-Net [16]	65.19%
E2E Learning [20]	68.71%
Local Relation Learning [19]	78.26%
MASDT [24]	80.21%
HFDCDNet (Proposed)	**84.26%**

cannot fully represent real-world conditions. Thus, detection models must not only perform well on known datasets but also maintain robustness when facing unseen data. To evaluate the generalization ability of the proposed HFDCD network, we conducted cross-dataset experiments by training the model on FaceForensics++ and testing it on Celeb-DF. The AUC metric was used to assess generalization performance, comparing against several representative detection methods. The results are summarized in Table 5, where the highest AUC values are highlighted in bold, indicating the best generalization ability among all compared approaches.

As shown in Table 5, the proposed HFDCD network demonstrates stronger adaptability when confronted with unseen data, outperforming the DCD network by more than two percentage points. This improvement mainly stems from the integration of high-frequency information and the design of the bidirectional cross-attention mechanism. The incorporation of high-frequency features enables the model to capture subtle details within images, preventing excessive reliance on coarse structures or general contours present in spatial features. Meanwhile, the bidirectional cross-attention effectively fuses spatial and frequency-domain features, ensuring the model neither over-relies solely on spatial nor frequency features, but rather leverages a balanced combination of both to guide classification decisions.

Table 6 Comparison of Network Parameters and Detection Accuracy

Methods	ACC	Parameters
F3-Net [16]	92.78%	57.3M
Xception [14]	95.04%	20.8M
MRT-Net [32]	95.50%	66.96M
AW-MSA [25]	96.65%	11.82M
HFDCD (Proposed)	98.76%	24.0M

With the rapid advancement of deepfake generation technologies, computational efficiency of detection models has become increasingly important. Although the proposed HFDCDNet exhibits excellent detection performance on public benchmark datasets, high-performing models often entail considerable computational complexity, which limits their practical deployment in resource-constrained environments. To comprehensively evaluate the practicality of HFDCDNet, we conducted computational resource experiments analyzing parameters such as model size, floating point operations (FLOPs), and inference speed. The results indicate that the HFDCD network requires 10.769 GFLOPs, contains 24.0 million parameters, and achieves an inference speed of 34.58 frames per second (FPS). Subsequently, we compared the parameter count and detection accuracy (on FaceForensics++) of HFDCDNet with several representative methods, as detailed in Table 6.

As shown in Table 6, the proposed HFDCD network achieves the highest detection accuracy while maintaining a parameter count of only 24 million, which is significantly lower than that of F3-Net and MET-Net. Although its parameter count is slightly higher than that of AW-MSA and Xception, its detection accuracy substantially surpasses both methods. Therefore, while AW-MSA–with its smaller network size and reasonable accuracy–may be more suitable for resource-constrained environments, the proposed HFDCD network trades a moderate increase in computational resources for a notable performance gain (2.11% improvement in accuracy). With a parameter count of just 24M, the HFDCD network demonstrates greater competitiveness in practical scenarios that balance computational capacity and demand for high detection accuracy.

4.3 Ablation study

To further validate the individual contributions of each functional component within the proposed HFDCD framework, we conducted a series of ablation experiments. Specifically, to assess the effectiveness of the DCD structure, we designed a simplified model incorporating only Deformable Convolutional Networks (DCN) and attention mechanisms. To separately evaluate the impact of DCN and attention modules, we further constructed models that include only DCN or only attention, respectively. In addition, we progressively

removed the bidirectional cross-attention module and the High-Frequency Enhancement (HFE) module to examine their contributions to the overall HFDCD architecture.

To verify the effectiveness of high-frequency preservation in the channel dimension within the HFE module, we implemented a variant that retains high-frequency information solely in the spatial dimension. Furthermore, to demonstrate the superiority of the proposed bidirectional cross-attention mechanism, we designed a comparative model utilizing simple concatenation and a unidirectional attention module.

All models were trained and tested on the FaceForensics++ dataset. The experimental results are summarized in Table 7, where the best-performing accuracy (ACC) and AUC values are highlighted in bold for clarity.

As shown in Table 7, both the Deformable Convolutional Network (DCN) and attention mechanisms contribute positively to enhancing the performance of deepfake detection. DCN significantly boosts model performance due to its superior spatial adaptability and its ability to recognize non-rigid deformations and irregular forged patterns. The attention mechanism also improves detection by enhancing key features, enabling the model to focus more effectively on critical regions of the input.

Although a simple combination of DCN and attention yields some performance improvement compared to using either component alone, its effectiveness is actually inferior. This could be attributed to the unstructured integration causing certain features to be overly emphasized, potentially leading to model overfitting. In contrast, the proposed DCD structure effectively addresses this issue, leading to a substantial performance gain.

After incorporating frequency-domain information, applying high-frequency retention only in the spatial dimension provides modest performance gains. However, this improvement is limited due to the lack of emphasis on channel-wise information. In contrast, our proposed High-Frequency Enhancement (HFE) module, which retains high-frequency components in both spatial and channel dimensions before frequency-domain learning, achieves significantly better results.

Table 7 Ablation experiment

Methods	ACC	AUC
Backbone	95.54%	98.99%
Backbone+DCN	96.45%	97.16%
Backbone+Attention	95.93%	96.33%
Backbone+DCN+Attention	95.59%	95.64%
Backbone+DCD	98.31%	99.51%
Backbone+DCD+HFE(Spatial FFT only)	98.38%	99.27%
Backbone+DCD+HFE	98.53%	99.33%
Backbone+DCD+HFE+(Concat&Attention)	98.53%	99.28%
Backbone+HFDCD	**98.94%**	**99.58%**

Moreover, a naive combination of concatenation and unidirectional attention introduces interference between spatial and frequency features, which hinders performance improvement. By contrast, the proposed HFDCD network leverages a bidirectional cross-attention mechanism to effectively fuse spatial and frequency-domain features. This selective enhancement and filtering of meaningful features from both domains results in optimal detection performance.

In addition, we conducted a comprehensive visual analysis using Class Activation Mapping (CAM) [42] to further interpret the ablation study results. This analysis reveals the differences in attention regions for various model configurations. Specifically, we computed CAMs on the final convolutional layer of the following five model variants: Backbone, Backbone+DCN, Backbone+Attention, Backbone+DCD, and Backbone+HFDCD. The resulting heatmaps were superimposed on the input images to visualize which regions were most influential in the model's decision-making process. The visualization results are presented in Fig. 10.

As shown in Fig. 10, the baseline backbone primarily relies on very small local regions for decision-making. This often leads to the neglect of other critical information within the image, thereby preventing the model from capturing high-level inconsistencies and ultimately impairing classification accuracy. The introduction of DCN and attention mechanisms effectively expands the model's focus area. However, simply incorporating DCN enables the model to extract broader features but may also cause it to focus on irrelevant or noisy regions. Similarly, using only an attention mechanism helps the model concentrate on facial areas, but its focus tends to be on a roughly rectangular region encompassing facial organs. This results in information loss at the boundary between the attended region and surrounding areas, reducing the model's ability to detect global features such as non-rigid deformations, facial coherence, forgery boundaries, and texture anomalies.

Moreover, as demonstrated in Table 7, the naive combination of DCN and attention leads to decreased accuracy. This may be due to feature conflicts, redundancy, or overfitting caused by their direct stacking. In contrast, the proposed Deformable Convolution and Attention-based Dual-branch (DCD) module achieves collaborative feature optimization through parallel processing, channel shuffling, and feature fusion. It effectively integrates the strong deformation modeling capability of DCN with the localized focus of attention mechanisms.

After incorporating the DCD module, the model can extract richer features while emphasizing the facial regions, thereby improving detection accuracy. The heatmaps produced by the proposed HFDCD network show a focus on facial keypoints (e.g., eyes, mouth, and nose), which are

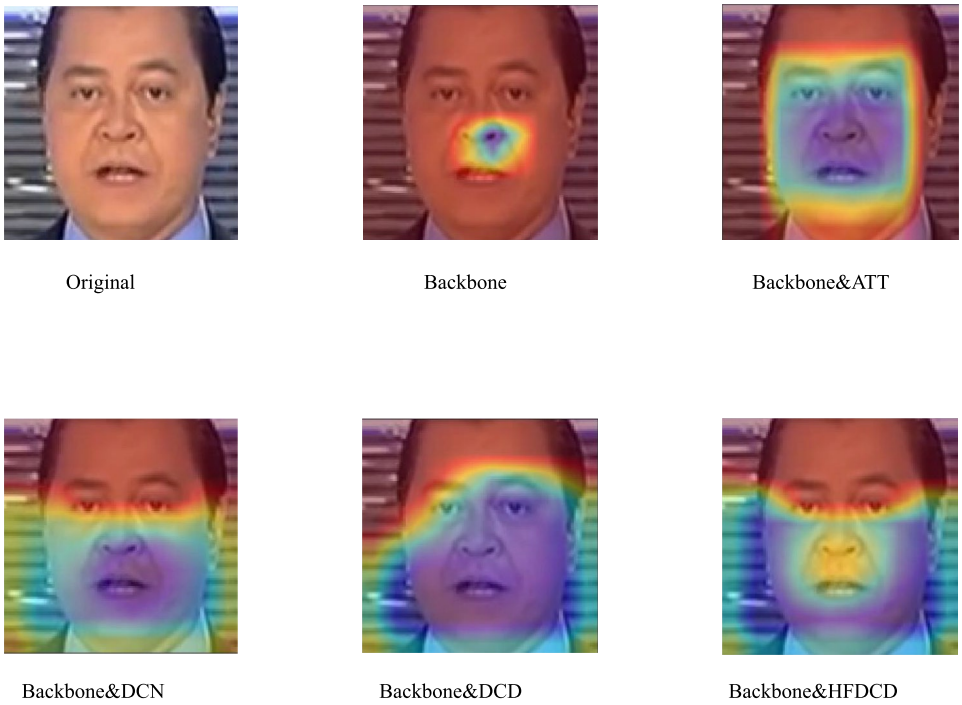


Fig. 10 Heatmap of different methods

typically indicative of forgery artifacts. This behavior aligns well with human intuition: humans tend to identify deepfake images by focusing on unnatural expressions or inconsistencies in facial details, such as mouth deformation or missing eye reflections. The HFE module further enhances this effect by capturing high-frequency features, which, although not easily perceived by the human eye, often correspond to subtle inconsistencies in forged regions. Finally, the bidirectional cross-attention mechanism fuses spatial and frequency-domain features, boosting the model's attention to these critical regions and enhancing detection performance. This comprehensive design leads to the superior detection capability demonstrated by the HFDCD network.

5 Conclusion

To address the limitations of traditional deepfake detection networks–such as insufficient feature extraction capabilities and inadequate attention to key facial regions–this paper proposes a spatial-domain feature extraction framework named the Deformable Convolution and Attention (DCD) module. The framework integrates two core components: a Deformable Convolutional Network (DCN) module, which enables more refined and comprehensive extraction of forgery cues, and an Attention module, which guides the network to focus more on critical facial areas. Moreover, recognizing the beneficial role of frequency-domain information in deepfake detection, we extend the DCD framework by incorporating frequency-domain analysis and feature fusion modules. This enables complementary enhancement of spatial and frequency representations, thereby further improving the model's detection capabilities.To validate the effectiveness of the proposed approach, we develop a novel network architecture, HFDCD, and conduct extensive experiments on the FaceForensics++ and Celeb-DF datasets. Comparative evaluations against several state-of-the-art methods demonstrate that our approach significantly improves classification performance by effectively leveraging both spatial and frequency features, outperforming many existing solutions. From ethical and societal perspectives, the proposed HFDCD framework holds substantial implications. Its high-accuracy detection capability can play a critical role in mitigating the malicious use of deepfake technologies in the spread of fake news, invasion of privacy, and defamation, thereby helping preserve the authenticity of digital information and public trust.

Despite its promising performance, the proposed method still has certain limitations, which point to important directions for future research. First, although it performs well on lightly compressed datasets, the model's robustness under significant quality degradation or strong noise interference remains underexplored. Second, this study focuses on image and video data and does not address the detection of forged audio, which represents a key limitation. Future work should aim to expand the detection capabilities to include audio-based manipulations.

Funding This work was supported in part by the basic research project of natural science of Yunnan province under Grant 202301AT070065, in part by the basic research project of education department of Yunnan

province under Grant 2023J0208 and 2023J0209.

Data Availability The datasets used in this study are publicly available.

Code Availability The code for this study will be released after the article is accepted for publication.

Declarations

Conflicts of Interest The authors have no financial or proprietary interests in any material discussed in this article.

References

1. Goodfellow I, Pouget-Abadie J, Mirza M et al (2014) Generative adversarial nets. Adv Neural Inf Process Syst 27
2. Mirza M, Osindero S (2014) Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784
3. Radford A, Metz L, Chintala S (2015) Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434
4. Brock A, Donahue J, Simonyan K (2018) Large scale GAN training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096
5. Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: Proceedings of IEEE/CVF conference on computer vision and pattern recognition, pp 4401–4410
6. Tolosana R, Vera-Rodriguez R, Fierrez J et al (2020) Deepfakes and beyond: a survey of face manipulation and fake detection. Inf Fusion 64:131–148
7. Thies J, Zollhofer M, Stamminger M et al (2016) Face2face: real-time face capture and reenactment of RGB videos. In: Proceedings on IEEE conference on computer vision and pattern recognition, pp 2387–2395
8. Thies J, Zollhöfer M, Nießner M (2019) Deferred neural rendering: image synthesis using neural textures. ACM Trans Graph (TOG) 38(4):1–12
9. Dosovitskiy A (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929
10. Chollet F (2017) Xception: deep learning with depthwise separable convolutions. In: Proceedings on IEEE conference on computer vision and pattern recognition, pp 1251–1258
11. Nguyen HH, Yamagishi J, Echizen I (2019) Use of a capsule network to detect fake images and videos. arXiv preprint arXiv:1910.12467
12. Dai J, Qi H, Xiong Y et al (2017) Deformable convolutional networks. In: Proceedings on IEEE international conference on computer vision, pp 764–773
13. Woo S, Park J, Lee JY et al (2018) CBAM: convolutional block attention module. In: Proceedings on European Conference on Computer Vision (ECCV), pp 3–19
14. Rossler A, Cozzolino D, Verdoliva L et al (2019) Faceforensics++: learning to detect manipulated facial images. In: Proceedings IEEE/CVF international conference on computer vision, pp 1–11
15. Afchar D, Nozick V, Yamagishi J et al (2018) MesoNet: a compact facial video forgery detection network. In: Proceedings IEEE Internationsl Workshop on Information Forensics and Security (WIFS), pp 1–7
16. Qian Y, Yin G, Sheng L et al (2020) Thinking in frequency: face forgery detection by mining frequency-aware clues. In: Proceedings on European conference on computer vision. Springer, Cham, pp 86–103
17. Nirkin Y, Wolf L, Keller Y et al (2021) Deepfake detection based on discrepancies between faces and their context. IEEE Trans Pattern Anal Mach Intell 44(10):6111–6121
18. Chen HS, Rouhsedaghat M, Ghani H et al (2021) Defakehop: a light-weight high-performance deepfake detector. In: Proceedings IEEE International Conference on Multimedia and Expo (ICME), pp 1–6
19. Chen S, Yao T, Chen Y et al (2021) Local relation learning for face forgery detection. In: Proceedings on AAAI conference on artificial intelligence, vol 35, no 2, pp 1081–1088
20. Cao J, Ma C, Yao T et al (2022) End-to-end reconstruction-classification learning for face forgery detection. In: Proceedings on IEEE/CVF conference on computer vision and pattern recognition, pp 4113–4122
21. Guo Z, Yang G, Zhang D et al (2023) Rethinking gradient operator for exposing AI-enabled face forgeries. Expert Syst Appl 215:119361
22. Lin H, Huang W, Luo W et al (2023) DeepFake detection with multi-scale convolution and vision transformer. Digital Signal Process 134:103895
23. Yang G, Wei A, Fang X et al (2023) FDS_2D: rethinking magnitude-phase features for DeepFake detection. Multimedia Syst 29(4):2399–2413
24. Das S, Kolahdouzi M, Özparlak L et al (2023) Unmasking Deepfakes: masked autoencoding spatiotemporal transformers for enhanced video forgery detection. In: Proceedings on IEEE International Joint Conference on Biometrics (IJCB), pp 1–11
25. Yadav A, Vishwakarma DK (2024) AW-MSA: adaptively weighted multi-scale attentional features for DeepFake detection. Eng Appl Artif Intell 127:107443
26. Tan C, Zhao Y, Wei S et al (2024) Frequency-aware deepfake detection: improving generalizability through frequency space domain learning. In: Proceedings on AAAI conference on artificial intelligence, vol 38, no 5, pp 5052–5060
27. Thao PNM, Dao CT, Wu C et al (2024) Medfuse: multimodal EHR data fusion with masked lab-test modeling and large language models. In: Proceedings 33rd ACM international conference on information and knowledge management, pp 3974–3978
28. Li Y, Yang X, Sun P et al (2020) Celeb-DF: a large-scale challenging dataset for deepfake forensics. In: Proceedings IEEE/CVF conference on computer vision and pattern recognition, pp 3207–3216
29. Shang Z, Xie H, Zha Z et al (2021) PRRNet: pixel-region relation network for face forgery detection. Pattern Recogn 116:107950
30. Xu K, Yang G, Fang X et al (2023) Facial depth forgery detection based on image gradient. Multimed Tools Appl 82(19):29501–29525
31. Yang Z, Liang J, Xu Y et al (2023) Masked relation learning for deepfake detection. IEEE Trans Inf Forensics Security 18:1696–1708
32. Yadav A, Vishwakarma DK (2023) MRT-Net: auto-adaptive weighting of manipulation residuals and texture clues for face manipulation detection. Expert Syst Appl 232:120898
33. Zhao H, Zhou W, Chen D et al (2021) Multi-attentional deepfake detection. In: Proceedings on IEEE/CVF conference on computer vision and pattern recognition, pp 2185–2194
34. Sun K, Liu H, Yao T et al (2022) An information theoretic approach for attention-driven face forgery detection. In: European conference on computer vision. Springer, Cham, pp 111–127
35. Khormali A, Yuan JS (2024) Self-supervised graph Transformer for deepfake detection. IEEE Access
36. Guo Z, Yang G, Wang D et al (2023) A data augmentation framework by mining structured features for fake face image detection. Comput Vis Image Underst 226:103587

37. Ke J, Wang L (2023) DF-UDetector: an effective method towards robust deepfake detection via feature restoration. Neural Netw 160:216–226
38. Choi DH, Lee HJ, Lee S et al (2020) Fake video detection with certainty-based attention network. In: Proceedings on IEEE International Conference on Image Processing (ICIP), pp 823–827
39. Guo Z, Yang G, Chen J et al (2021) Fake face detection via adaptive manipulation traces extraction network. Comput Vis Image Underst 204:103170
40. Chen Z, Yang H (2021) Attentive semantic exploring for manipulated face detection. In: Proceedings on IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 1985–1989
41. Wang J, Wu Z, Ouyang W et al (2022) M2TR: multi-modal multi-scale transformers for deepfake detection. In: Proceedings on international conference on multimedia retrieval, pp 615–623
42. Zhou B, Khosla A, Lapedriza A et al (2016) Learning deep features for discriminative localization. In: Proceedings on IEEE conference on computer vision and pattern recognition, pp 2921–2929