DeepFake Detection Based on High-Frequency Enhancement Network for Highly Compressed Content

Jie Gao[a,b], Zhaoqiang Xia[a], Gian Luca Marcialis[b], Chen Dang[a,b], Jing Dai[c] and Xiaoyi Feng[a,d,*]

[a] *School of Electronics and Information, Northwestern Polytechnical University, 1 Dongxiang Road, Xi'an, 710129, Shaanxi, China*
[b] *Department of Electrical and Electronic Engineering, University of Cagliari, Piazza d'Armi, Cagliari, 09123, Sardegna, Italy*
[c] *China Academy of Launch Vehicle Technology, No. 1 Nan Da Hong Men Road, Beijing, 100076, China*
[d] *Shenzhen Research Institute of Northwestern Polytechnical University, No. 45 Gaoxin South 9th Road, Shenzhen, 518057, China*

ARTICLE INFO

Keywords:
Low-Quality DeepFake
DeepFake Detection
High-Frequency Enhancement
Discrete Cosine Transform
Discrete Wavelet Transform

ABSTRACT

The DeepFake, which generates synthetic content, has sparked a revolution in the fight against deception and forgery. However, most existing DeepFake detection methods mainly focus on improving detection performance with high-quality data while ignoring low-quality synthetic content that suffers from high compression. To address this issue, we propose a novel High-Frequency Enhancement framework, which leverages a learnable adaptive high-frequency enhancement network to enrich weak high-frequency information in compressed content without uncompressed data supervision. The framework consists of three branches, i.e., the Basic branch with RGB domain, the Local High-Frequency Enhancement branch with Block-wise Discrete Cosine Transform, and the Global High-Frequency Enhancement branch with Multi-level Discrete Wavelet Transform. Among them, the local branch utilizes the Discrete Cosine Transform coefficient and channel attention mechanism to indirectly achieve adaptive frequency-aware multi-spatial attention, while the global branch supplements the high-frequency information by extracting coarse-to-fine multi-scale high-frequency cues and cascade-residual-based multi-level fusion by Discrete Wavelet Transform coefficients. In addition, we design a Two-Stage Cross-Fusion module to effectively integrate all information, thereby greatly enhancing weak high-frequency information in low-quality data. Experimental results on FaceForensics++, Celeb-DF, and OpenForensics datasets show that the proposed method outperforms the existing state-of-the-art methods and can effectively improve the detection performance of DeepFakes, especially on low-quality data. The code is available here [1].

1. Introduction

In recent years, with the development of social media, the Internet has provided a massive amount of information, which not only makes it easier for humans to obtain knowledge, but also makes it possible for misleading information, fake news, and other synthetic content to spread. The synthesis of fake content [19] has aroused great interest because it promotes progress in certain fields, such as the use of digital people to replace real people in the field of live broadcasts, the interaction between the virtual and the real world, the different expressions and actions of characters in film and television production, etc. Nonetheless, it is difficult to ignore the huge harm it causes [37], especially the fake facial content known as DeepFake [62, 28]. Through facial replacement or expression manipulation, malicious attackers create fake news videos, making viewers mistakenly believe that the fictional events actually happened, which further raises public concerns about the credibility of information and privacy security. To counter this technology, an arms race has begun in the development of face forgery detectors [55, 2, 26].

To date, a considerable amount of DeepFake detection methods [70, 19] have been developed. For instance, a series of works [55, 11, 4] have explored the existence of low-level artifact traces. Shiohara and Yamasaki [55] constructed self-blended images (SBIs) to detect DeepFakes by reproducing common forgery artifacts. Chen *et al.* [4] designed a Bi-granularity artifacts detector to explore the intrinsic-granularity artifacts caused by model

[1] https://github.com/Gina-YUE/HiFE

*Corresponding author.

jie_gao@mail.nwpu.edu.cn (J. Gao); zxia@nwpu.edu.cn (Z. Xia); marcialis@unica.it (G.L. Marcialis); cdang@mail.nwpu.edu.cn (C. Dang); buaadaij@126.com (J. Dai); fengxiao@nwpu.edu.cn (X. Feng)

ORCID(s): 0000-0003-2823-2682 (J. Gao); 0000-0003-0630-3339 (Z. Xia); 0000-0002-8719-9643 (G.L. Marcialis); 0000-0003-3389-5413 (C. Dang); 0000-0002-3130-6884 (J. Dai); 0000-0002-0428-6224 (X. Feng)

[1] DOI: https://doi.org/10.1016/j.eswa.2024.123732

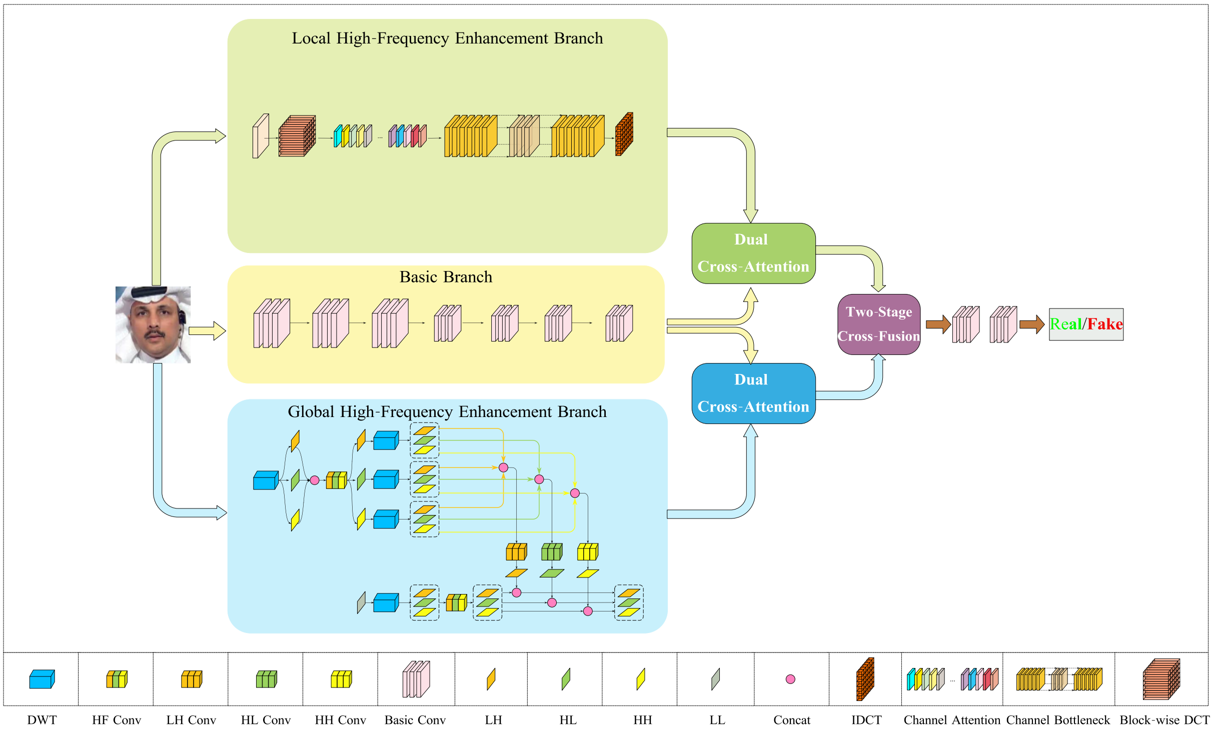


Figure 1: Structural diagram of the proposed framework. It consists of three branches: (1) Basic branch, which uses original RGB information; (2) adaptive Local High-Frequency Enhancement branch based on Block-wise DCT, channel attention mechanism, and inverse Block-wise DCT; (3) adaptive Global High-Frequency Enhancement branch based on Multi-level DWT and cascade-residual-based multi-level fusion strategy. Furthermore, the Two-Stage Cross-Fusion module is designed to achieve complementary information from three different domains.

generation and extrinsic-granularity artifacts due to blend operation. Dong *et al.* [11] proposed an ID-unaware model that aims to remove the influence of facial identity information, thereby using only artifact clues for authentic and fake identification. Furthermore, some methods based on mid-level manipulation traces are devoted to finding inconsistencies [14, 36, 71, 69] in fake content. In [14], the inconsistency of identity information between inner and outer faces is explored. Liu *et al.* [36] developed a detector that focuses on temporal identity inconsistency. Yu *et al.* [71] proposed an augmented multi-scale spatiotemporal inconsistency magnifier to capture comprehensive synthesis spatiotemporal clues. Yang *et al.* [69] distinguished real from fake by detecting inconsistent 3D head poses caused by manipulation. Moreover, some biological signals [17, 50, 65, 16] are also applied to this task. Hernandez-Ortega *et al.* [17] developed a detector named DeepFakesON-Phys, which uses heart-rate estimation to judge real and fake videos. Qi *et al.* [50] developed DeepRhythm, which captures differences in heart rate variations between genuine and fake videos by detecting subtle changes in skin color caused by facial blood flow. Wu *et al.* [65] regarded DeepFake detection as a source detection task and utilized the multi-scale spatiotemporal photoplethysmography map from multiple facial regions to capture counterfeit cues. In [16], the author performed inter- and intra-modality reasoning on emotions extracted from audio and visual modalities for DeepFake detection. However, these efforts primarily focus on the outstanding performance of high-quality, uncompressed content, lacking analysis of key factors leading to performance degradation when encountering low-quality fake content.

Considering that spatial artifacts and synthetic traces often suffer degradation during the compression process, making it more challenging to detect compressed data, some methods utilize frequency domain information to mitigate this issue, such as Fast Fourier Transform (FFT) [12, 38], Discrete Cosine Transform (DCT) [10, 51, 5], and Discrete Wavelet Transform (DWT) [45]. Nevertheless, most of these methods use single-frequency transform. Although these studies can address those above fragile spatial artifacts to some extent, the cues extracted by FFT and manually

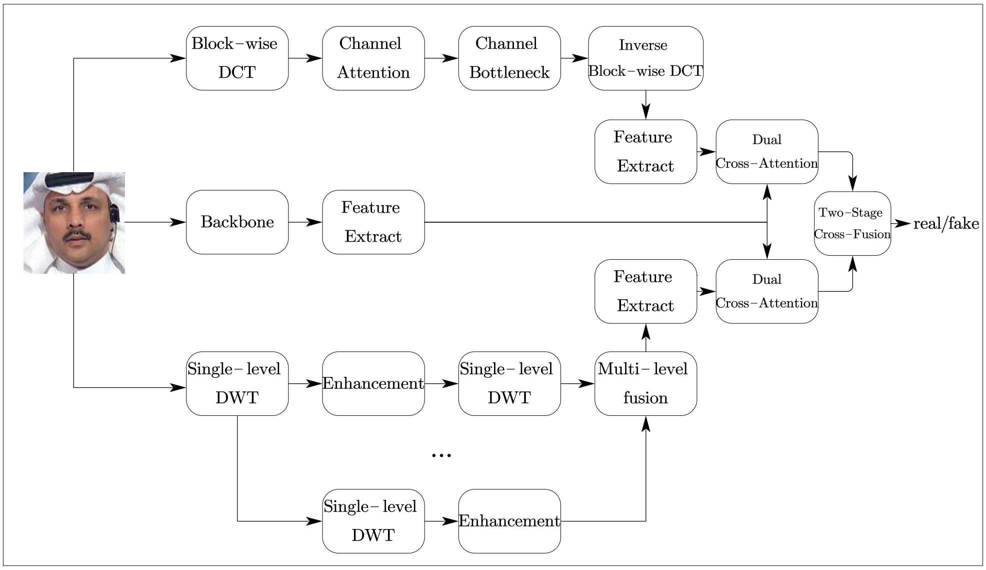


Figure 2: A running example of the proposed method is used to demonstrate the entire model running pipeline. The input image is fed into three branches for feature extraction in different domains, and final fusion is performed for model inference and prediction.

partitioned DCT multi-frequency information only provide limited global frequency-domain information. On the other hand, analyzing the low and high-frequency components extracted by DWT ignores the local spatial relationship. These previous works motivate us to explore the complementarity between the frequency and wavelet domains.

In this work, we focus on improving the detection performance of high-compression DeepFake data. Considering that obtaining high-quality uncompressed data is often challenging in practical application scenarios, even if it is available, deploying supervised learning based on uncompressed high-quality data will incur high computational costs. Therefore, in this paper, we adopt an unsupervised learning strategy without high-quality data. Our goal is to bridge the detection performance gap between high-quality and low-quality data by enhancing the weakened high-frequency information in low-quality data through an unsupervised learning strategy.

Before constructing the model, we conducted extensive experimental analysis as described in Section 3.1. We perform frequency information decomposition tests on both uncompressed data and their corresponding highly compressed data. Moreover, we design three strategies to verify the importance of high-frequency information in low-quality DeepFake discrimination. Based on the above massive analysis, we propose a new architecture named High-Frequency Enhancement (**HiFE**), as shown in Figure 1.

Specifically, we employ Discrete Cosine and Wavelet Transforms to adaptively enhance the residual local and global high-frequency information in low-quality data. Furthermore, we explore their complementarity and design a fusion module. The strength of the proposed method lies in the adaptive utilization of high-frequency information in low-quality data through local Discrete Cosine Transform and global Discrete Wavelet Transform, without the need for supervision from high-quality data. Additionally, our module can be flexibly embedded into any model with extremely low computational burden. The main contributions of this work are summarized as follows:

- We analyze the difference between low-quality and high-quality data through quantitative and qualitative experiments, revealing the key factors leading to good detection performance on high-quality data. In addition, we designed three strategies of information removal, information exchange, and information re-learning through discrete wavelet decomposition to verify the conjecture further quantitatively.

- We propose a High-Frequency Enhancement (HiFE) network to handle low-quality data. To the best of our knowledge, this is the first method that utilizes adaptive local and global high-frequency enhancement branches to explore the most favorable high-frequency information for compressed DeepFake detection without the need for uncompressed content supervision. Additionally, we design a Two-Stage Cross-Fusion strategy to simulate the complementarity between different domains.

- The adaptive frequency-aware features from local sub-networks adopt Block-wise DCT, channel attention mechanism, channel bottleneck module, and inverse Block-wise DCT. Moreover, Multi-level DWT decomposition layers and cascade-residual-based multi-level fusion strategies are designed to realize adaptive global high-frequency enhancement.

- Extensive comparative and thorough ablation experiments demonstrate that the proposed method achieves significant improvements and state-of-the-art performance on low-quality DeepFake data with a small computational burden.

The remaining parts of the paper are organized as follows: Section 2 introduces the standard DeepFake generation approaches and their corresponding detection methods. In Section 3, we provide a detailed explanation of our proposed framework and its intuition. The experimental details and settings are presented in Section 4. Finally, in Section 5, we draw conclusions based on the experimental results, summarize the advantages and limitations of the proposed method, and discuss the prospect.

2. Related Work

This section focuses on DeepFake generation principles and a series of detection methods, especially the most relevant work to solve the compression problem, including related applications of DCT and DWT.

2.1. DeepFake Generation Methods

As one of the most widely discussed technologies, DeepFake originated from the Reddit website in 2017 [46]. A user uploaded a malicious fake video on the website, garring significant public attention and sparking extensive discussions. In recent years, with the continuous development of artificial intelligence technology, various DeepFake synthesis technologies have been proposed, and some applications have been developed, such as FaceSwap [2], DeepFaceLab [3], ZAO [4], ReFace [5], and so on. These advancements in deep learning technology have enabled anyone without specialized knowledge to synthesize forged content. For instance, current technology can already create people who do not exist, or directly change the speaker's lip shape and expression in a real video to easily create a corresponding fake video, which was unimaginable before. The increasingly realistic synthetic content is flooding the entire Internet, leading to people's contemplation and concern about the credibility of information.

DeepFake technology can automatically map the facial expressions or identity ID of a source person onto a target person. With appropriate post-processing, the generated video can have a remarkably realistic visual appearance. Considering that this technology is easily abused by criminals, a large number of DeepFake synthetic video datasets have been publicly used for detector development research. Among them, the most widely used methods for automatic face manipulation and forgery, including Deepfakes [24], Face2Face [58], FaceSwap [25], NeuralTextures [59], and FaceShifter [30] which can be found in FaceForensics++ (FF++) [54] dataset. Advanced DeepFake generation algorithms can successfully deceive and trick the human eye. Especially with the continuous development of DeepFake generation techniques, the recognition of synthetically generated fake data becomes increasingly challenging. For instance, datasets such as Celeb-DF [32] and OpenForensics [27], generated using improved synthesis methods, significantly enhance the visual quality of fake images. This renders detection methods that performed well on previous datasets inadequate in addressing these new challenges.

There are many types of research exploring facial forgery techniques. Among them, the two most commonly used and effective forgery architectures are Variational Auto-Encoder (VAE) [23] and Generative Adversarial Networks (GAN) [15]. The GAN-based methods have been continuously driving the sustainable development of this field, consistently improving the visual realism of generated images. The basic principle behind them is a pair of generator/discriminator, where the generator network aims to generate fake images that are indistinguishable from real images, and the discriminator network aims to distinguish them. After repeated games, the generator can finally synthesize images with realistic appearances. The strategy of this kind of adversarial game compels the generator to effectively outmaneuver the discriminator, which is the fundamental reason why forgery models can generate high-quality manipulated image content [60]. Especially in recent years, various improvements based on GANs have made

[2] https://github.com/MarekKowalski/FaceSwap/, (Last accessed: 18.02.2024)

[3] https://github.com/iperov/DeepFaceLab/, (Last accessed: 18.02.2024)

[4] https://zaodownload.com/,(Last accessed: 18.02.2024)

[5] https://reface.app/,(Last accessed: 18.02.2024)

it possible to generate partially or completely fake media. For example, Singh *et al.* [56] utilized StyleGAN2 [22] to produce synthetic facial images with variable facial expressions. Prajwal *et al.* [49] designed a Wav2Lip model based on GAN, aiming to improve the lip-sync accuracy of synthesis video. Liu *et al.* [39] proposed a face-swapping model by combining the attention mechanism and CycleGAN to alleviate the discontinuities across the blending boundaries. Tao *et al.* [57] achieved high-quality text-to-image synthesis with the help of GAN. Kang *et al.* [21] presented a new open-source library named StudioGAN, which supports many types of GAN architecture for image synthesis.

2.2. DeepFake Detection Methods

The continuous advancement of DeepFake synthesis methods has also propelled the development of corresponding detection techniques. Currently, DeepFake detection has gained significant attention and plays a crucial role in ensuring the security of facial recognition systems. Unfortunately, according to numerous references, most existing DeepFake detection methods primarily focus on improving the performance of high-quality data (i.e., data with low compression ratios) and achieving incredible performance for these data, while ignoring the challenges of low-quality data caused by high compression ratios [18]. Although significant progress has been made in the forensic detection of high-resolution image/video datasets, the prevalence of low-quality DeepFake data in practical scenarios caused by data compression during transmission calls for further exploration and research. Therefore, we will introduce detection methods specifically designed for high-quality data and methods dedicated to low-quality data separately below.

Methods for High-Quality Data. A large number of methods are dedicated to addressing the detection of high-quality forged data and have demonstrated promising performance. These methods are generally based on three categories of observations: low-level visual artifacts, mid-level manipulation traces, and high-level semantic information.

Low-level. Li *et al.* [29] used the inconsistency between different color channels (e.g. HSV or YCbCr) to determine whether the image is forged. Matern *et al.* [43] pointed out that the facial features of fake face images have some inconsistencies of visual details (eyes, nose bridge, teeth, etc.), which can be used for fake face detection. Li and Lyu [34] determined whether it is a fake video by detecting whether the video contain the distortion caused by the affine transformation of scaling, rotating, and cutting. Shiohara and Yamasaki [55] constructed self-blended images (SBIs) to detect DeepFakes by reproducing common forgery artifacts, such as blended boundaries and statistical inconsistencies between source and target images. Zhou *et al.* [75] designed a fake face detection framework based on two streams, one is GoogLeNet to detect tampering artifacts in a face classification stream, and another is a patch triplet stream to capture local noise residuals and camera characteristics.

Mid-level. Wang *et al.* [61] proposed Fakespotter, which extracts the activation state of neurons for real and fake detection. Yang *et al.* [69] considered inconsistencies in 3D head pose after synthesized face region stitching. According to [31], fake faces typically involve a fusion step. Based on this assumption, they proposed Face X-ray to detect the presence of face fusion boundaries in images and use it as a metric to measure the authenticity. Afchar *et al.* [1] constructed a mid-level semantic classifier called MesoNet because they believed that the differences between genuine and fake content are more evident at the intermediate semantic level.

High-level. In [17], DeepFakesON-Phys was developed, which uses heart rate estimation to detect fake content. Qi *et al.* [50] constructed DeepRhythm to capture differences in heart rate variations between genuine and fake videos, and they incorporated a motion-amplified spatiotemporal representation module and a dual spatiotemporal attention module to better adapt to dynamically changing faces and various forgery types. Ciftci *et al.* [7] utilized physiological signals from facial regions for real and fake distinguish, the spatial and temporal coherences are computed using transformations, and signal features are captured from the PPG graph. Mittal *et al.* [47] introduced a Siamese network-based approach to extract and analyze the similarity between audio and visual modalities in the same video. By simultaneously extracting emotional cues from both modalities and constructing a triple loss, they aim to capture the differences between genuine and fake videos.

However, the methods mentioned above are based on the assumption of uncompressed or lightly compressed data. These methods assume that fake data will not be heavily compressed and that manipulation traces can be preserved. Unfortunately, once forged content is highly compressed using codecs like JPEG or H.264, the forged traces are susceptible to compression artifacts, making detection more challenging.

Methods for Low-Quality Data. Recognizing the limitations of the above approaches, several studies have observed and explored the issue from different perspectives. Although there are relatively few studies in this area, we roughly divide them into three categories: DCT-based, DWT-based, and Others.

DCT-based. Given the advantages of fast operation speed and energy concentration, Discrete Cosine Transform is often used not only in the field of data compression but also in other tasks. For instance, Qian *et al.* [51] introduced frequency into face forgery detection by applying DCT as the applied frequency-domain transformation. Concas *et al.* [10] proposed a novel feature extraction approach by exploiting the DCT domain to deal with the problem of compressed images. In order to increase the amount of information collected by small images, they considered overlapping block operations and further utilized the statistics of DCT coefficients to describe the input data.

DWT-based. Wavelet transforms can separate data into spatial frequency components of different directions and scales. Due to their lossless and reversible nature, they are commonly used in various fields such as super-resolution, image reconstruction, and image denoising [53, 66]. However, research on DWT has not been fully exploited in DeepFake detection tasks. As far as we know, it has only been explored in [45]. Miao *et al.* [45] improved the ability to capture general fine-grained artifacts in the frequency domain by exploiting DWT decomposition and discarding low-frequency components.

Others. In addition, some works are dedicated to exploring spatial information. Hu *et al.* [18] designed a dual-stream network for detecting compressed DeepFake videos, where the least compressed I-frames are used during the training of the frame-level stream instead of the entire frame. Obviously, this method is generally not suitable for highly compressed fake images. Liao *et al.* [35] constructed a facial-muscle-motions-based (FAMM) framework to solve the problem of compressed DeepFake video detection, but this cannot be used for compressed images. Moreover, Li *et al.* [33] proposed a Spatial Restoration Detection Framework (SRDF). Inspired by super-resolution techniques, they designed feature extraction enhancement and mapping modules with the help of uncompressed videos to restore and enhance the texture features of low-quality compressed videos. Woo *et al.* [64] proposed an Attention-based DeepFake Detection Distiller (ADD) by utilizing knowledge distillation (KD), where they designed a distillation process to transfer the distribution of teacher tensors from high-quality (HQ) data to student tensors called low-quality (LQ) data, achieving good performance. However, their approach relies on the assumption that high-quality images are easily obtainable, which is often challenging in practice.

In contrast to the above methods, our proposed approach simultaneously utilizes both DCT and DWT transforms for enhancing local and global high-frequency information without supervision from uncompressed high-quality data. Compared to the other methods, our proposed method is more generalizable and flexible as it can be embedded into any network.

2.3. Unsupervised and Fusion Strategies

Except for the above-mentioned works in 2.2, there are also some kinds of literature devoted to the research of unsupervised learning and fusion strategies.

Compared to supervised learning, unsupervised learning does not require large amounts of labeled training data, reducing the cost and workload associated with data annotation. Additionally, through unsupervised learning, models are less prone to getting stuck in local optima, making it easier for the algorithm to adapt to the overall structure of the data rather than specific labels. Therefore, this strategy has been widely applied across multiple domains. Currently, there are also some studies exploring its application in DeepFake detection tasks. For example, Fung *et al.* [13] used data augmentation methods to generate two different versions of an image and achieved unsupervised training by encouraging the model to learn the maximum similarity between the two images. Qiao *et al.* [52] designed a pseudo-label generator to label training samples and fed them into the proposed enhanced contrastive learner for binary classification of synthetic videos. Mejri *et al.* [44] introduced a framework named UNTAG, which is based on unsupervised learning, for type-agnostic DeepFake detection to alleviate inadequate generalization.

Additionally, the concept of fusion strategy has brought new insights to the task of DeepFake detection. In [9], the authors exploit the complementarity of different individual classifiers and evaluate which fusion rules are best suited to improve the generalization capabilities of modern DeepFake detection systems. Omar *et al.* [48] proposed a deep learning bagging ensemble classifier to detect manipulated faces in videos. They combined the predictions from multiple learners to classify the video as real or fake. Zhang *et al.* [72] built Particle Swarm Optimization (PSO) based weighted and evolving ensemble models integrating optimized 3D Convolutional Neural Networks and Recurrent Neural Networks for video authenticity classification.

The success of the above strategies motivates us to explore more efficient and convenient detection models.

3. The Proposed Method

In this section, we start by analyzing the differences between compressed DeepFake content and its corresponding uncompressed versions, highlighting the motivation behind designing a high-frequency information enhancement network from the frequency domain perspective. Additionally, we conduct comparative experiments to demonstrate the importance of high-frequency information in detecting low-quality synthesized content. Subsequently, we provide a more detailed explanation of the proposed method and architecture.

3.1. Motivation and Analysis

In this part, we analyze the motivation for using high-frequency enhancement strategies and a series of verification experiments. Specifically, we conduct experimental verification on three different compressed versions of the FaceForensics++ [54] dataset. As shown in Figure 3 and Figure 4, where "raw" denotes the uncompressed high-quality version (HQ), "c23" represents a lightly compressed medium-quality version (MQ), and "c40" indicates a highly compressed low-quality version (LQ).

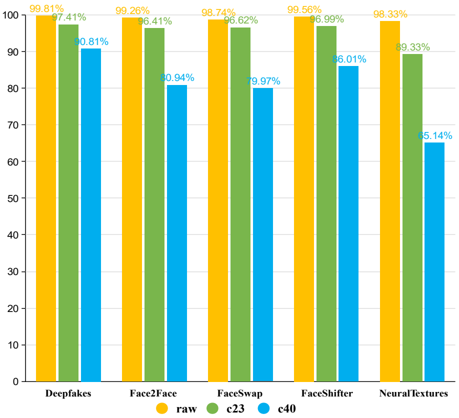
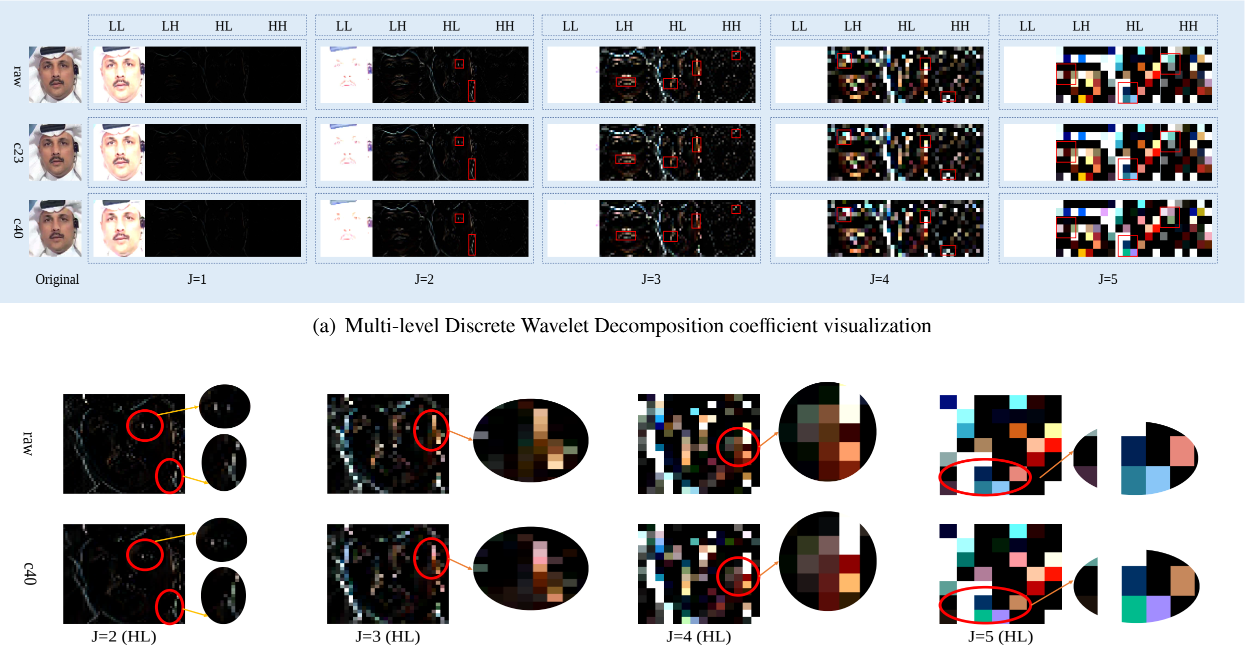


Figure 3: Performance comparison of Xception network on aligned FF++ [54] dataset with different compression ratios. Among them, raw is the uncompressed high-quality (HQ) version, c23 refers to the weakly compressed medium-quality (MQ) version, and c40 represents the highly compressed low-quality (LQ) version.

Quantitative analysis. (*The performance degradation of low-quality data.*) DeepFake detection methods based on deep learning have made significant progress in evaluating high-quality data. However, the performance often degrades dramatically when testing with low-quality data that suffer from highly compressed. As shown in Figure 3, We perform experiments on aligned data with different compression ratios, and experimental studies show that the stronger the compression, the worse the performance of the DeepFake detection model. Considering that these data are the same except for the compression, we believe that the reason for this phenomenon is the loss of information during the compression process, which essentially discards high-frequency information that does not visible to the human eye while preserving low-frequency information. Additionally, the quantization process introduces compression artifacts, greatly affecting the detection of forged trace details.

Qualitative analysis. (*The difference between the visualization of low-quality data and high-quality data.*) In order to find the difference between uncompressed data and aligned compressed data more intuitively, we perform visual analysis by using discrete wavelet decomposition. Discrete Wavelet Transform is a traditional image decomposition method that can decompose a two-dimensional image into a combination of low-frequency and high-frequency images in multiple directions while preserving the spatial information of the image. Therefore, the structural information and detailed information of the original image can be easily extracted by using this transformation. Based on the above advantages, we decided to use wavelet decomposition to analyze the detection performance to observe which part is more conducive to the improvement of DeepFake detection performance. As shown in Figure 4 (a), we perform multi-level discrete wavelet decomposition visual analysis on the same image of three different compression versions, where “J” represents the level of wavelet decomposition, “LL” represents low-frequency information, and “LH”, “HL” and “HH” represent high-frequency information in different directions, respectively. From Figure 4 (a), we can easily find that with the increase in decomposition levels, the difference between low-quality and high-quality data becomes



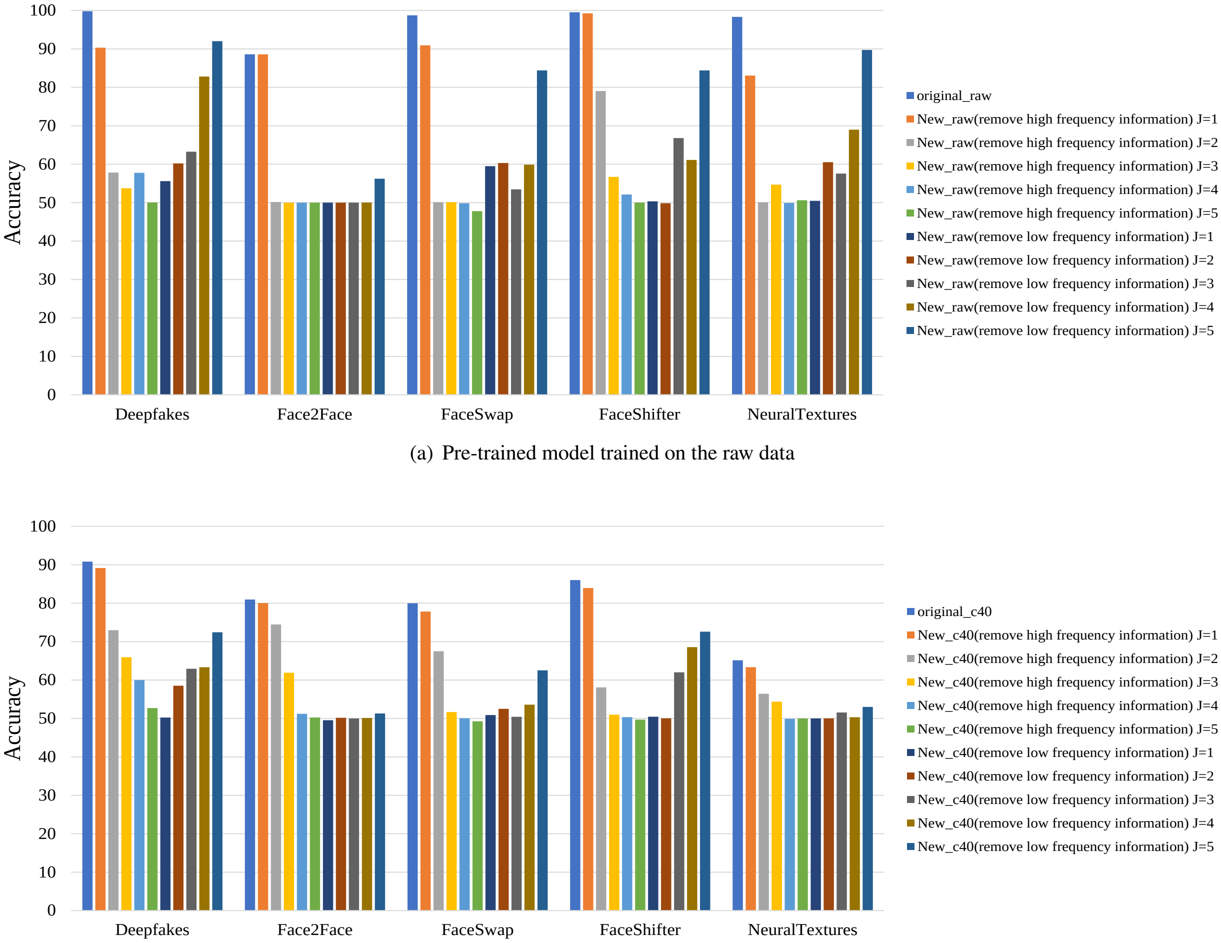
(a) Multi-level Discrete Wavelet Decomposition coefficient visualization

(b) The Comparison of HL high-frequency components between uncompressed (raw) and highly-compressed (c40) data

Figure 4: Illustration of using multi-level discrete wavelet decomposition. LL represents low-frequency components; LH, HL, and HH represent high-frequency information in horizontal, vertical, and diagonal directions. The J represents the decomposition level. The red line area is the decomposition visualization result of the aligned data under different compression strengths. Figure (a) shows the results for all decomposed components. For a clearer observation, Figure (b) only shows the comparison results of the HL component.

increasingly obvious, especially in high-frequency components. The higher the compression ratio, the more obvious the difference. In order to illustrate the difference between highly compressed and uncompressed data in high-frequency information more clearly, we take "HL" as an example in Figure 4 (b) for further visual analysis. Obviously, the high-frequency coefficients are greatly changed after compression.

Information removal. In order to further analyze the impact of high-frequency information on DeepFake detection performance, we design the information removal strategy. Specifically, we replace the high-frequency coefficients obtained through discrete wavelet decomposition with 0 and then apply the inverse transformation to restore the signal after removing the high-frequency information. Similarly, we set the low-frequency coefficients to 0 to eliminate low-frequency information. Then we use the pre-trained model to test new combined data. As shown in Figure 5 (a), we conducted experiments on uncompressed raw data. Specifically, we employed two strategies: high-frequency information removal and low-frequency information removal. Under each strategy, we use the pre-trained model on raw data to test new composite data. From the results of Figure 5 (a), we can observe that the detection accuracy decreases as the high-frequency information is gradually removed. Conversely, the accuracy improves as the low-frequency information is removed to a lesser extent, and correspondingly more high-frequency information is retained. We can better illustrate this point with two intuitive examples. For example, when removing high-frequency information, let's consider the case where $J = 1$. The performance starts to decline in Figure 5 (a). As we increase J to 5, the performance becomes the worst. On the contrary, let us consider the opposite scenario. When performing remove low-frequency information with $J = 5$, the low-frequency information is removed the least, while the corresponding high-frequency information is kept the most. The accuracy rate is the highest at this time, compared with $J = 1$ to $J = 4$. Similarly, as a further comparative illustration, we conducted the same experiments on highly compressed data (c40) [54], as shown in Figure 5 (b). It is evident that we arrive at a consistent conclusion that the more high-frequency information is retained, the more advantageous it is for DeepFake detection tasks.

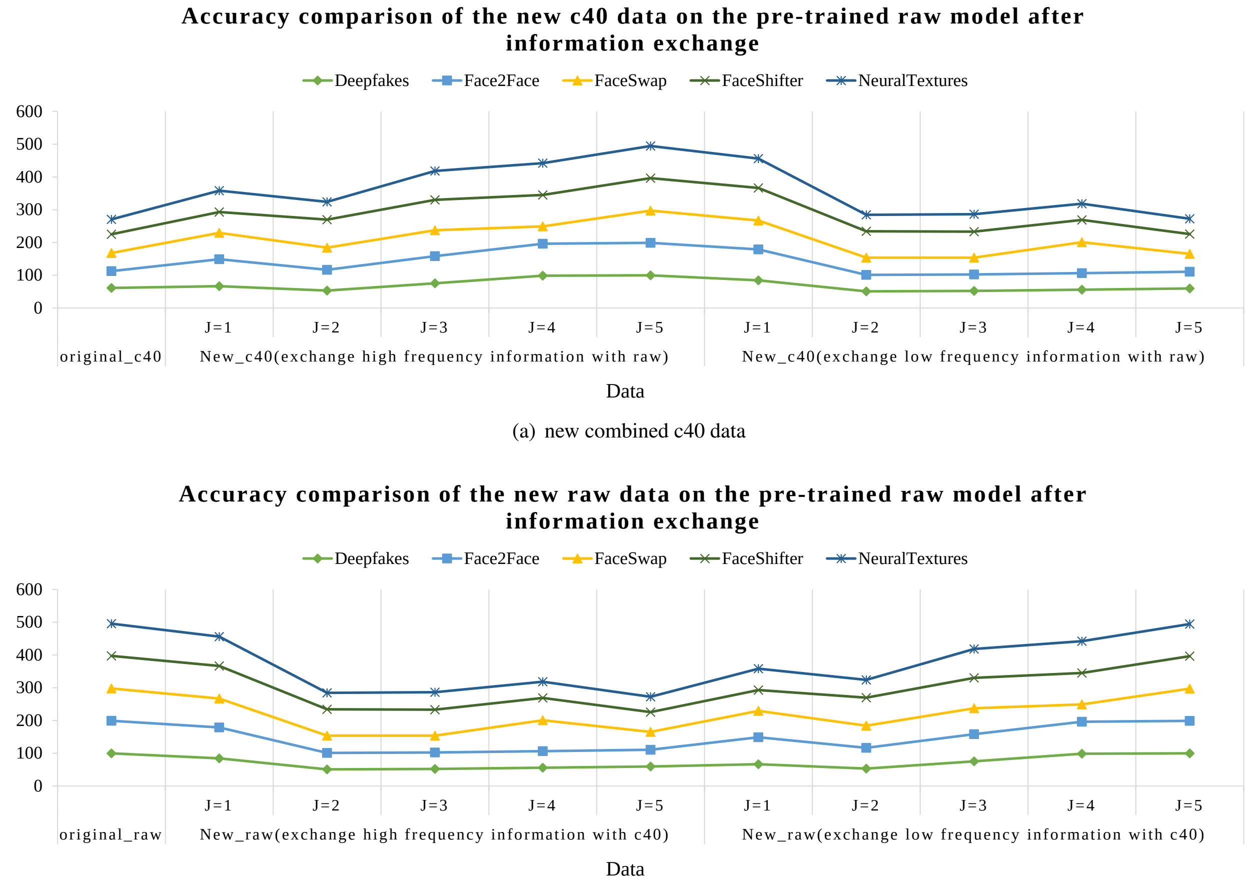


(a) Pre-trained model trained on the raw data

(b) Pre-trained model trained on the c40 data

Figure 5: Comparative validation experiments based on removing different frequency domain information. Figure (a) illustrates the detection performance of the pre-trained model, which is trained on the uncompressed version data (raw), on newly synthesized data obtained by applying high-frequency and low-frequency removal. Similarly, Figure (b) presents the experimental results of the pre-trained model trained on the highly compressed version data (c40), showcasing the model's performance on newly synthesized data obtained by applying high-frequency removal and low-frequency removal on the compressed c40 data.

Information exchange. Furthermore, we utilize discrete wavelet decomposition to perform information exchange strategies between the uncompressed raw data and c40 data. We aim to explore the differences in information learned by deep neural networks between high-quality and low-quality data, and further validate the importance of high-frequency information in DeepFake detection tasks. To achieve this goal, we design two information exchange strategies: high-frequency information exchange and low-frequency information exchange. As shown in Figure 6, we perform information exchange between the c40 data and the raw data to synthesize new data (new c40 data and new raw data). Additionally, we utilize the pre-trained model on the original raw data to test the new synthesized data. In Figure 6, "original_c40/original_raw" represents the original data (c40 or raw), and "J" represents the wavelet decomposition level. For example, in Figure 6 (a), "New_c40 (exchanging high-frequency information with raw), $J = 1$" means that we replace the first-level high-frequency information of the original c40 data with the first-level high-frequency information from the original raw data. To illustrate this operation more intuitively, we visualize the new synthesized data after information exchange, as shown in Figure 7. We visualize comparing information exchange results for wavelet decomposition levels $J = 1$ to $J = 5$. For high-frequency information exchange, a higher decomposition



(a) new combined c40 data

(b) new combined raw data

Figure 6: Comparative verification experiment based on frequency domain information exchange strategy. In Figure (a), we perform high-frequency information exchange and low-frequency information exchange on the c40 and aligned raw data. Then, we utilize the pre-trained model based on raw data to test the newly synthesized c40 data. Similarly, in Figure (b), we test the newly synthesized raw data with the pre-trained model based on the raw data.

level means more high-frequency information is exchanged, resulting in the newly combined c40 data containing more uncompressed high-frequency information that comes from raw data. On the contrary, the quality of newly synthesized raw data will be decreased. In addition, the experimental results in Figure 6 indicate that the richer the high-frequency information, the higher the corresponding accuracy in DeepFake detection. This phenomenon suggests that the promising detection performance of deep neural networks on uncompressed data is mainly attributed to the learning of high-frequency information. However, for low-quality data, deep neural networks can only rely on low-frequency information for learning, resulting in a sharp decline in detection performance. This observation indirectly reflects that CNN-based deepfake detection methods mainly rely on high-frequency information (Texture) in synthetic data for learning, which is consistent with the research results of Liu *et al.* [40].

Information re-learning. Moreover, we retrain a new model on the newly synthesized c40 data after the information exchange to explore whether the information enhancement strategy can improve detection accuracy. As shown in Table 1, we analyze the results after exchanging different levels of high-frequency information. It can be found that when less low-frequency information is retained, more high-frequency information is obtained through the exchange (e.g. $J = 5$), and the synthesized data is more easily identified. In other words, the high-frequency enhancement strategy is beneficial to the authenticity of low-quality data.

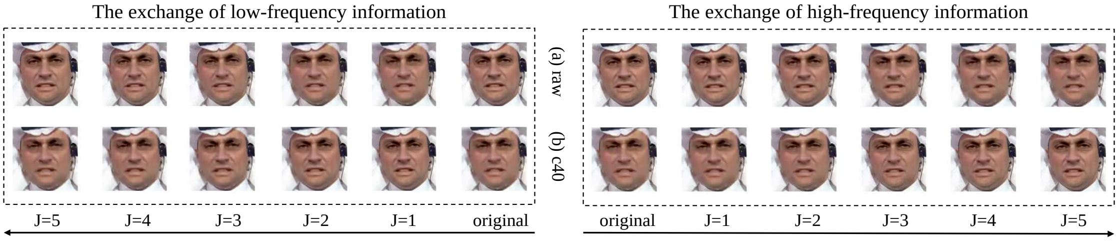

The exchange of low-frequency information
The exchange of high-frequency information
(a) raw
(b) c40
J=5
J=4
J=3
J=2
J=1
original
original
J=1
J=2
J=3
J=4
J=5

Figure 7: New synthetic data results based on information exchange. The left sub-figure shows the result obtained by low-frequency information exchange, and the right sub-figure shows the result obtained by high-frequency information exchange. The first row, "(a) raw", represents the visualization process of the newly synthesized raw data as information swapping takes place. The second row, "(b) c40", represents the visualization process of the newly combined c40 data as information exchange occurs. Obviously, in the left sub-figure, as "J" gets larger, the c40 data becomes blurrier, while the original data becomes clearer. On the right sub-figure, the raw data becomes more blurred as "J" gets larger, while the c40 data behaves in the opposite direction.

Table 1
The performance of the model that retrained on the newly synthesized c40 data, where the high-frequency information in the synthesized c40 data is derived from the original uncompressed raw data.

Data type	original c40	New c40 data				
		J=1	J=2	J=3	J=4	J=5
DeepFakes	**90.81**	99.69	99.83	99.83	99.93	***99.97***
Face2Face	**80.94**	99.48	99.44	99.59	99.59	***99.61***
FaceSwap	**79.97**	98.01	98.54	99.13	99.16	***99.34***
FaceShifter	**86.01**	98.63	98.99	99.49	99.53	***99.55***
NeuralTextures	**65.14**	98.06	98.78	99.10	99.26	***99.33***

3.2. High-Frequency Enhancement Network

Based on a series of analyses above, we propose a novel **High-Frequency Enhancement (HiFE)** network for highly compressed low-quality fake content from a novel perspective. Our goal is to solve the problem of poor performance for low-quality data. In particular, HiFE aims to explore the most effective information for the DeepFake detection task, and the architecture is shown in Figure 1. The proposed detection framework consists of three branches, i.e., the Basic branch based on original RGB, the Local High-Frequency Enhancement (LHiFE) branch based on Block-wise DCT, and the Global High-Frequency Enhancement (GHiFE) branch based on Multi-level DWT.

On the one hand, to reflect the importance of frequency domain information in different local spatial domains, we construct a novel frequency-aware multi-spatial attention strategy based on Block-wise DCT. With the learnable channel attention mechanism and channel bottleneck module, we are able to adaptively extract and enhance local high-frequency information. In addition, the channel bottleneck module can also remove redundant information and improve the utilization of effective information. Finally, we restore the new frequency-aware coefficients to the spatial domain via inverse Block-wise DCT.

On the other hand, to make full use of multi-directional and multi-scale high-frequency information, we design a global high-frequency enhancement branch through Multi-level DWT embedding. Specifically, we construct a Multi-level DWT and a cascade-residual-based multi-level fusion strategy to explore and enhance the weakened global high-frequency information in low-quality data.

Finally, we design a Two-Stage Cross-Fusion module for the multi-domain information fusion, which is built after Dual Cross-Attention to effectively embed local high-frequency and global high-frequency information into the backbone network.

To provide a clearer illustration of how the input image propagates through the model during the forward pass, we present a running example (Figure 2) as a supplement to describe the entire model pipeline of our proposed architecture.

3.2.1. Local High-Frequency Enhancement Branch with Block-wise DCT

Considering that the importance of various local regions is different, we aim to construct a Local High-Frequency Enhancement branch that fully utilizes the frequency domain information to implement an adaptively frequency-aware multi-spatial attention mechanism in different local spatial domains.

To achieve the above goal, we analyze existing compression techniques and point out that compression affects different local regions differently. Moreover, we approach the problem from a novel perspective by leveraging the principles of compression in reverse.

Image compression is a common digital image processing strategy to reduce transmission pressure. In particular, JPEG technology divides the image into non-overlapping image blocks, performs discrete cosine transformation on each image block to obtain DCT coefficients, and finally quantizes coefficients through a quantization table, which will cause high-frequency loss. While this operation effectively reduces storage pressure overall, quantization has different effects on different local regions of the image. For instance, certain local regions may lose more high-frequency information compared to other regions. Losing some high-frequency information may not disrupt the content of the image, but it can be detrimental to neural networks. Studies [68, 41, 73] have shown that in the process of target recognition, the neural network will first use low-frequency information, and after reaching a certain level, it will deeply dig high-frequency information to improve accuracy. This phenomenon also confirms why compressed images are harder to discern. In order to fully exploit the importance of information in different local areas of the image, we reproduce the DCT compression process. The difference is that our purpose is to extract the local high-frequency information and give more attention to the remaining high-frequency components rather than eliminating high-frequency details. The specific algorithm is shown in Algorithm 1.

```
Algorithm 1 Algorithm in the Local High-Frequency Enhancement branch (LHiFE)
Require: RGB image $X \in \mathbb{R}^{C\times H\times W}$                              ▷ Input
Ensure: Local High-Frequency Enhancement image $X_{IBD} \in \mathbb{R}^{C\times H\times W}$   ▷ Output
  1: $X \to \{X_i, X_i \in \mathbb{R}^{C\times h\times w}\}$                                   ▷ Block image
  2: $X \to X_{BD} \in \mathbb{R}^{64*C\times M\times N}$                                      ▷ Block-wise DCT, Eq. (1)
  3: $X_{BD} \in \mathbb{R}^{64*C\times M\times N} \to X_{CA} \in \mathbb{R}^{64*C\times M\times N}$   ▷ Channel Attention, Eq. (2),(3)
  4: $X_{CA} \in \mathbb{R}^{64*C\times M\times N} \to X_{CB} \in \mathbb{R}^{64*C\times M\times N}$   ▷ Channel Bottleneck, Eq. (4)
  5: $X_{CB} \in \mathbb{R}^{64*C\times M\times N} \to X_{IBD} \in \mathbb{R}^{C\times H\times W}$     ▷ Inverse Block-wise DCT, Eq. (5)
```

To illustrate the proposed method more clearly, we represent the local high-frequency branch graphically. As shown in Figure 8, suppose the input image is $X \in \mathbb{R}^{C\times H\times W}$. Firstly, we partition the image X into non-overlapping blocks X_i of size $h * w$. Based on that operation, we get a total of $M * N$ sub-blocks, where $M = \frac{H}{h}, N = \frac{W}{w}$. The input image can be represented as $X = \{X_i\}_{i=1}^{M*N}$, and $X_i \in \mathbb{R}^{C\times h\times w}$. Then, we transform the spatial domain information into the frequency domain X_{BD} by performing the Block-wise Discrete Cosine Transform (BD), which can be expressed by Eq. (1):

$$X_{BD} = \{D(X_i)\}_{i=1}^{M*N} \tag{1}$$

Where $D(\cdot)$ indicates the DCT operation. After the above operation, all the DCT coefficients are spliced together to form multiple DCT coefficient channels. Since we perform the 8*8 blocking operation on the image, we can get 64 DCT coefficients, and X_{BD} satisfies $X_{BD} \in \mathbb{R}^{64*C\times M\times N}$.

Furthermore, we adopt the channel attention mechanism (CA) to adaptively realize information interaction and high-frequency enhancement of different DCT coefficient channels. As shown in Eq. (2), with the channel attention mechanism, we adaptively adjust the importance of different DCT coefficient channels and achieve relative enhancement of high-frequency information. After that, we obtain new channel weights $W_{CA} \in \mathbb{R}^{64*C\times 1\times 1}$ through the sigmoid function.

$$W_{CA} = sigmoid\left(MLP\left(AvgPool\left(X_{BD}\right)\right) + MLP\left(MaxPool\left(X_{BD}\right)\right)\right) \tag{2}$$

By reweighting the original DCT coefficient channel, the new DCT coefficient channel $X_{CA} \in \mathbb{R}^{64*C\times M\times N}$ can be obtained, as shown in Eq. (3), where the symbol $\odot$ represents the element dot product operation.

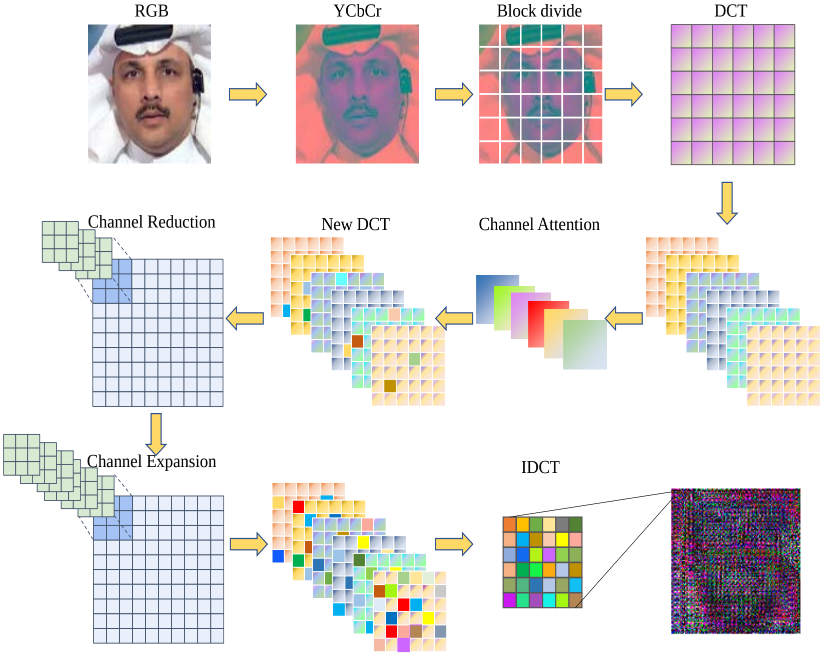


Figure 8: The proposed Local High-Frequency Enhancement branch (LHiFE) pipeline based on Block-wise DCT.

$$X_{CA} = X_{BD} \odot W_{CA} \tag{3}$$

In addition, we construct a learnable channel bottleneck (CB) module represented by the convolution module F_{CB}, which is designed to remove redundant information and further achieve adaptive fine-tuning enhancement of DCT high-frequency coefficients. Moreover, by using multiple 3*3 convolution stacks, this module achieves information interaction between adjacent spatial locations in the same DCT channel, and the local field of view of the high-frequency enhancement strategy is expanded, thereby effectively retaining spatial information correlation. After that, the new channel coefficients $X_{CB} \in \mathbb{R}^{64*C\times M\times N}$ can be obtained. The operation can be referred to Eq. (4), where $\otimes$ represents convolution operation:

$$X_{CB} = F_{CB} \otimes X_{CA} \tag{4}$$

Finally, the DCT coefficients of different channels adaptively weighted by the channel attention mechanism and the channel bottleneck module are back to the spatial domain $X_{IBD} \in \mathbb{R}^{C\times H\times W}$ through Inverse Block-wise Discrete Cosine Transform (IBD), as depicted in Eq. (5).

$$X_{IBD} = \left\{ D^{-1}(X_{CB}) \right\} \tag{5}$$

Through the above operations, the New DCT coefficients that highlight high-frequency information will be reflected in different local spatial domains, thereby indirectly realizing the multi-spatial attention mechanism of frequency awareness. Additionally, the designed algorithm can directly reflect the frequency domain information differences of genuine and fake content, which are difficult for neural networks to capture, into different local spatial domain images. Experimental experience tells us that this not only helps to make full use of frequency domain information but also improves the training efficiency of the model.

3.2.2. *Global High-Frequency Enhancement Branch with Multi-level DWT*

In order to augment the global image texture representation with high-frequency information for robust high-compression image detection, we propose a Global High-Frequency Enhancement branch. In particular, we employ discrete wavelet decomposition to decompose the image into multiple sub-bands of high-frequency and low-frequency information to avoid destroying the spatial structure. Furthermore, we construct a multi-level enhancement strategy for the high-frequency information obtained by different decomposition levels, aiming to exploit synthetic high-frequency trajectories deeply.

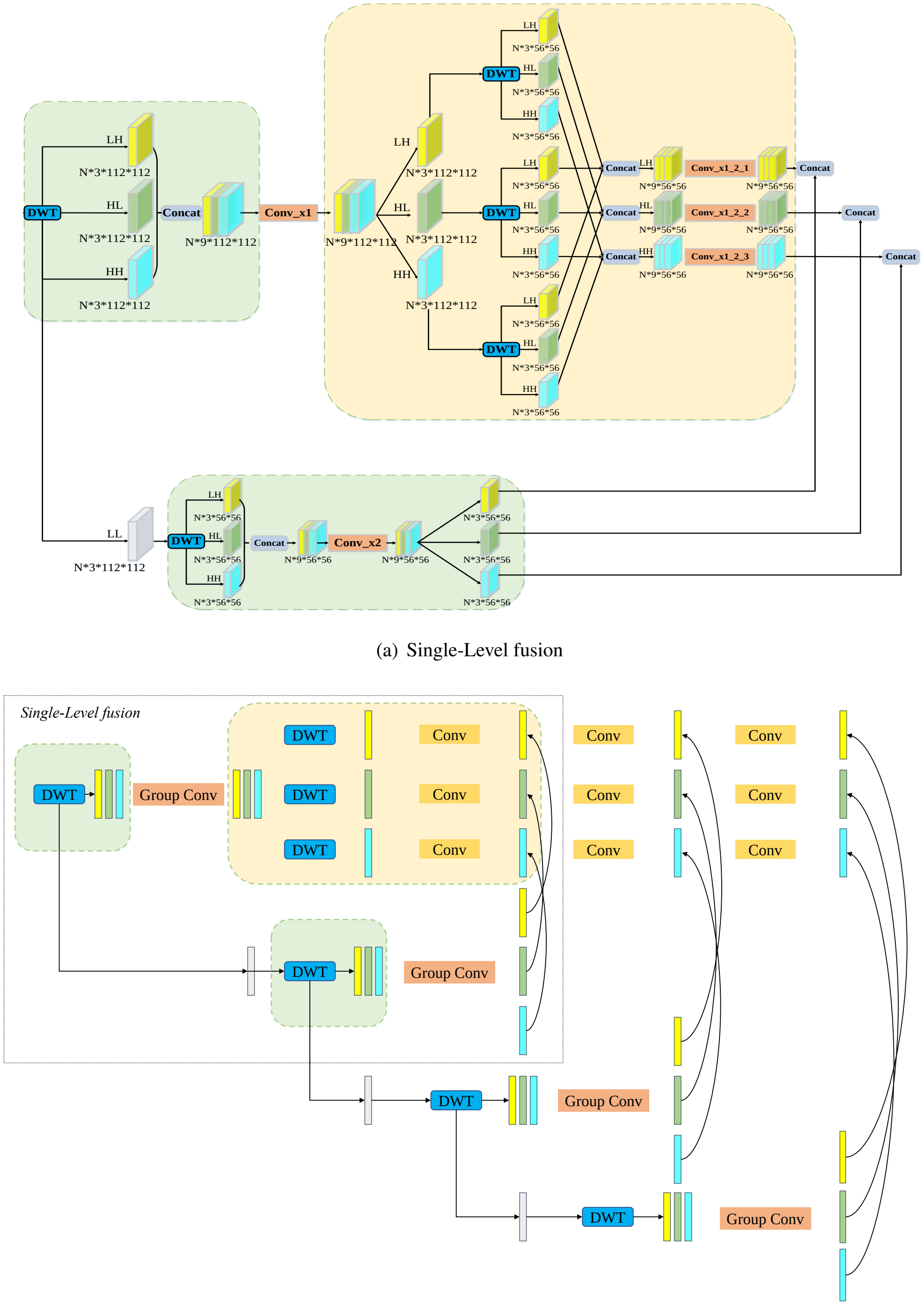
As shown in Figure 9 (b), we design a cascade-residual-based multi-level fusion strategy, which is dedicated to deeply mining high-frequency information in low-quality images. The high-frequency information from the previous level is further decomposed into three components (horizontal information, vertical information, diagonal information), and they are cascaded with the corresponding three types of high-frequency information from the next level. The purpose of this operation is to gradually supplement the residual high-frequency information and prevent the loss of weak high-frequency information during the forward propagation process of the deep neural network. In addition, after each level of DWT, we constructed the corresponding high-frequency enhancement module to reinforce high-frequency information at different scales adaptively. Moreover, to describe our algorithm more clearly, we provide a detailed single-level fusion strategy in Figure 9 (a). And the overall global high-frequency enhancement algorithm is shown in Algorithm 2.

Algorithm 2 Algorithm in the Global High-Frequency Enhancement branch (GHiFE)

```
Require: RGB image $X \in \mathbb{R}^{C\times H\times W}$, The total number $J$ of decomposition levels    ▷ Input
Ensure: Global High-Frequency Enhancement image $X_{MLD} \in \mathbb{R}^{C\times\frac{H}{16}\times\frac{W}{16}}$    ▷ Output
for $i \leftarrow 1$ to $J$ do
    $X \rightarrow \left\{X_{i,LL}, X_{i,LH}, X_{i,HL}, X_{i,HH}\right\}, X_{i,*} \in \mathbb{R}^{C\times\frac{H}{2^i}\times\frac{W}{2^i}}$    ▷ Single-Level DWT, Eq. (7,10)
    $X_{i,LH}, X_{i,HL}, X_{i,HH} \leftarrow F_{HF} \otimes Concat\left(X_{i,LH}, X_{i,HL}, X_{i,HH}\right)$    ▷ Eq. (8,11)
    if $i > 1$ then
        $\begin{cases} X_{i,LH} \leftarrow F_{LH} \otimes Concat\left(X_{i,LH}, X_{LH}^{i-1}\right) \\ X_{i,HL} \leftarrow F_{HL} \otimes Concat\left(X_{i,HL}, X_{HL}^{i-1}\right) \\ X_{i,HH} \leftarrow F_{HH} \otimes Concat\left(X_{i,HH}, X_{HH}^{i-1}\right) \end{cases} \quad X_{i,*} \in \mathbb{R}^{C\times\frac{H}{2^i}\times\frac{W}{2^i}}$    ▷ Multi-Level Fusion, Eq. (12)
        $X_{i,H} = Concat\left(X_{i,LH}, X_{i,HL}, X_{i,HH}\right)$    ▷ Enhanced Global High-Frequency Information
    else if $i = J$ then
        $X_{MLD} \leftarrow F_{ConvBlock} \otimes Concat\left(X_{i,LH}, X_{i,HL}, X_{i,HH}\right)$    ▷ Final Fusion Results
    end if
    $\begin{cases} X_{i,LH} \rightarrow \left\{X_{i,LL}^{LH}, X_{i,LH}^{LH}, X_{i,HL}^{LH}, X_{i,HH}^{LH}\right\}, X_{i,*}^{LH} \in \mathbb{R}^{C\times\frac{H}{2(i+1)}\times\frac{W}{2(i+1)}} \\ X_{i,HL} \rightarrow \left\{X_{i,LL}^{HL}, X_{i,LH}^{HL}, X_{i,HL}^{HL}, X_{i,HH}^{HL}\right\}, X_{i,*}^{HL} \in \mathbb{R}^{C\times\frac{H}{2(i+1)}\times\frac{W}{2(i+1)}} \\ X_{i,HH} \rightarrow \left\{X_{i,LL}^{HH}, X_{i,LH}^{HH}, X_{i,HL}^{HH}, X_{i,HH}^{HH}\right\}, X_{i,*}^{HH} \in \mathbb{R}^{C\times\frac{H}{2(i+1)}\times\frac{W}{2(i+1)}} \end{cases}$    ▷ Multi-Level DWT
    $\begin{cases} X_{LH}^{i} \leftarrow F_{LH} \otimes Concat\left(X_{i,LH}^{LH}, X_{i,LH}^{HL}, X_{i,LH}^{HH}\right) \\ X_{HL}^{i} \leftarrow F_{HL} \otimes Concat\left(X_{i,HL}^{LH}, X_{i,HL}^{HL}, X_{i,HL}^{HH}\right) \\ X_{HH}^{i} \leftarrow F_{HH} \otimes Concat\left(X_{i,HH}^{LH}, X_{i,HH}^{HL}, X_{i,HH}^{HH}\right) \end{cases} \quad X_{*}^{i} \in \mathbb{R}^{3C\times\frac{H}{2(i+1)}\times\frac{W}{2(i+1)}}$    ▷ Multi-Direction, Eq. (9)
    $X \leftarrow X_{i,LL}$    ▷ Multi-Scale: Decomposition of Low-Frequency Information
end for
```

We use the following notation for a detailed description of the first-level high-frequency information enhancement learning in Algorithm 2. Similar to Section 3.2.1, assuming that the input image can be represented by $X \in \mathbb{R}^{C\times H\times W}$, we adopt the multi-level wavelet decomposition strategy to globally separate frequency domain information. As shown in Eq. (6), there are four wavelet decomposition filters in each stage, denoted as f_{LL}, f_{LH}, f_{HL}, and f_{HH}, respectively.

$$\begin{cases} f_{LL} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} & f_{LH} = \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix} \\ f_{HL} = \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix} & f_{HH} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \end{cases} \tag{6}$$



(a) Single-Level fusion

(b) Multi-Level fusion

Figure 9: Schematic diagram of the Global High-Frequency Enhancement (GHiFE) branch based on Multi-level DWT. Figure (a) describes the single-level fusion strategy in detail; Figure (b) shows the multi-level fusion strategy, and the part inside the dotted rectangle box represents a rough representation of the single-level fusion.

Decomposed by the first-level DWT, we can obtain one low-frequency information $X_{1,LL}$ and three high-frequency information $X_{1,LH}$, $X_{1,HL}$, $X_{1,HH}$ with different directions, which can be represented by Eq. (7). Where $\otimes$ represents convolution operation, and $\downarrow_2$ refers to the image size being reduced by half.

$$\begin{cases} X_{1,LL} = (f_{LL} \otimes X) \downarrow_2 \\ X_{1,LH} = (f_{LH} \otimes X) \downarrow_2 \\ X_{1,HL} = (f_{HL} \otimes X) \downarrow_2 \\ X_{1,HH} = (f_{HH} \otimes X) \downarrow_2 \end{cases} \tag{7}$$

After that, we perform adaptive learning on the decomposed high-frequency information through Eq. (8) to achieve single-level high-frequency information-enhanced learning. Among them, F_{HF} refers to our three-directional high-frequency enhancement module, which is mainly composed of enhancement module F_E and group learning module $F_{GroupConv}$. The purpose of group learning is to ensure the separation of information in different directions for the next step of fusion. The F_E module utilizes multiple convolution stacks with a kernel size of 1*1 to achieve enhancement purposes. For the convenience of description, we uniformly use F_E to refer to the enhancement module in the following and provide the composition of this module in Figure 10.

$$X_{1,LH}, X_{1,HL}, X_{1,HH} = F_{HF} \otimes Concat\left(X_{1,LH}, X_{1,HL}, X_{1,HH}\right) \tag{8}$$

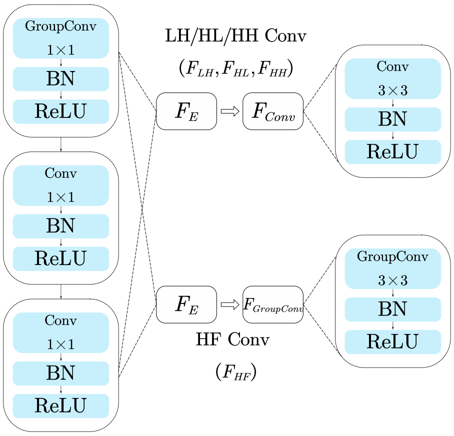


Figure 10: The network architecture of some components (HF/HL/LH/HH Conv in Fig. 1) in the GHiFE branch.

Additionally, to obtain multi-scale high-frequency information, we further decompose the first-level high-frequency information learned above. We perform DWT operations separately on the high-frequency information in three different directions and concatenate them accordingly. Subsequently, we apply enhancement strategies individually in each direction, as described in Eq. (9). Here we use different symbols (F_{LH},F_{HL},F_{HH}) to represent high-frequency information enhancement learning modules in different directions.

$$\begin{cases} X^1_{LH} = F_{LH} \otimes Concat\left\{\left(f_{LH} \otimes X_{1,LH}\right) \downarrow_2, \left(f_{LH} \otimes X_{1,HL}\right) \downarrow_2, \left(f_{LH} \otimes X_{1,HH}\right) \downarrow_2\right\} \\ X^1_{HL} = F_{HL} \otimes Concat\left\{\left(f_{HL} \otimes X_{1,LH}\right) \downarrow_2, \left(f_{HL} \otimes X_{1,HL}\right) \downarrow_2, \left(f_{HL} \otimes X_{1,HH}\right) \downarrow_2\right\} \\ X^1_{HH} = F_{HH} \otimes Concat\left\{\left(f_{HH} \otimes X_{1,LH}\right) \downarrow_2, \left(f_{HH} \otimes X_{1,HL}\right) \downarrow_2, \left(f_{HH} \otimes X_{1,HH}\right) \downarrow_2\right\} \end{cases} \tag{9}$$

At the same time, we excavate high-frequency information in different directions from the first-level low-frequency information to supplement the above-mentioned insufficient high-frequency information, which can be described as

Eq. (10).

$$\begin{cases} X_{2,LL} = (f_{LL} \otimes X_{1,LL}) \downarrow_2 \\ X_{2,LH} = (f_{LH} \otimes X_{1,LL}) \downarrow_2 \\ X_{2,HL} = (f_{HL} \otimes X_{1,LL}) \downarrow_2 \\ X_{2,HH} = (f_{HH} \otimes X_{1,LL}) \downarrow_2 \end{cases} \tag{10}$$

Likewise, we use a three-directional high-frequency enhancement module F_{HF} to perform feature extraction and reinforcement on them, as shown in Eq. (11).

$$X_{2,LH}, X_{2,HL}, X_{2,HH} = F_{HF} \otimes Concat\left(X_{2,LH}, X_{2,HL}, X_{2,HH}\right) \tag{11}$$

Furthermore, we design a fusion strategy based on cascaded residuals. This strategy is similar to residual, but the difference is that we only perform cascade fusion of the high-frequency information obtained by decomposing the low-frequency information of the previous layer and the high-frequency information obtained by re-decomposing the previous layer, instead of simple addition. Specifically, according to the direction, we combined the high-frequency information (X^1_{LH}, X^1_{HL}, X^1_{HH}), which is obtained by re-decomposing and learning from the first-level high-frequency information, and the high-frequency information ($X_{2,LH}$, $X_{2,HL}$, $X_{2,HH}$), which is collected by decomposing and learning from the first-level low-frequency information.

$$X_{1,H} = Concat\begin{Bmatrix} F_{LH} \otimes Concat\left(X_{2,LH}, X^1_{LH}\right) \\ F_{HL} \otimes Concat\left(X_{2,HL}, X^1_{HL}\right) \\ F_{HH} \otimes Concat\left(X_{2,HH}, X^1_{HH}\right) \end{Bmatrix} \tag{12}$$

Finally, by cascading high-frequency information in different directions and further reinforcing strategy, the enhanced first-level global high-frequency information $X_{1,H}$ is attained, which can be expressed by Eq. (12). So far, we have achieved first-level global high-frequency information fusion results.

In order to obtain multi-level global high-frequency information, we continue to decompose and learn the obtained high-frequency information $X_{1,H}$ and low-frequency information $X_{2,LL}$ similar to the first-level above. After multiple decompositions and fusions, we obtain the multi-level global high-frequency components $X_{4,H}$ (we set J=4 in Algorithm 2), and then we construct a feature extract block $F_{ConvBlock}$ to get the final Multi-level Discrete Wavelet Transform high-frequence feature X_{MLD}.

In our work, the proposed global branch based on the Discrete Wavelet Transform employs a four-level wavelet decomposition to ensure compatibility between the output size of high-frequency components and the backbone network, while also extracting more high-frequency information. In other words, following the same process, we continue to obtain three levels of global high-frequency enhancement information, allowing the 224x224 input image to be decomposed into 14x14, which is then fused with the features of the backbone network.

3.2.3. Two-Stage Cross-Fusion Module

Considering the complementarity of local and global information, we construct a Two-Stage Cross-Fusion module for multi-domain information fusion, aiming to further improve the model's robustness in the face of compression. In particular, we perform the fusion of the Local High-Frequency Enhancement branch with the Global High-Frequency Enhancement at the mid-stream position of the backbone network. On the one hand, choosing this location guarantees feature size compatibility. On the other hand, the mid-stream position of the backbone network can not only extract enough RGB image features, but the remaining backbone network can fully and effectively fuse the three-domain information after this position.

In addition, before the fusion module we designed, our single high-frequency enhancement branch and backbone network first adopted the Dual Cross-Attention (DCA) mechanism [42]. We hope that the module can adaptively pay different levels of attention to the information coming from the two domains. As shown in Figure 1, on the one hand, we use DCA to perform feature re-adjustment between the LHiFE branch and the backbone network. On the other hand, we perform feature adaptive fine-tuning between the GHiFE branch and the backbone network in the same way. After that, we embed the designed Two-Stage Cross-Fusion module to further learn and fuse the above features.

By designing a two-stage fusion method, we not only avoid the conflicts caused by the direct fusion of global high-frequency information and local frequency perception information but also effectively retain all available information, and can learn adaptively in the second stage of fusion, making the learned high-frequency information is more fully integrated into the backbone network. The schematic diagram of our fusion strategy is shown in Figure 11.

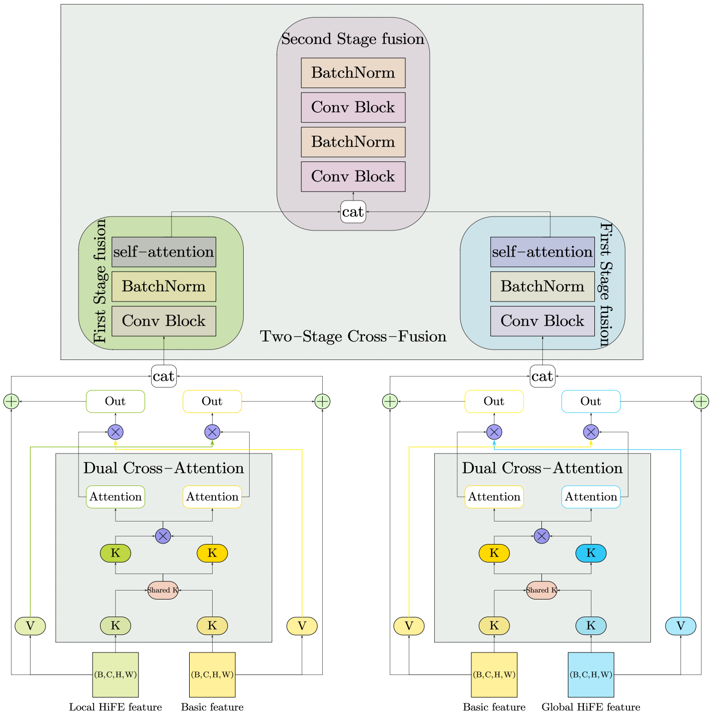


Figure 11: Schematic diagram of the Dual Cross-Attention and Two-Stage Cross-Fusion module.

4. Experiments

4.1. Experimental Setup

Datasets. In order to get a more complete and reliable conclusion, following most of the previous works on face forgery detection, we choose to conduct experiments on the challenging FaceForensics++ [54] dataset which contains three compression levels and five forgery methods, and Celeb-DF [32] dataset that is more realistic. Moreover, we also choose the OpenForensics [27] dataset that is dedicated to DeepFake detection in the wild as a supplement to highlight the superiority of our method.

FaceForensics++. FF++ is a face forgery detection video dataset containing 1000 real videos and 5000 fake videos. To synthesize fake images, Faceforensics++ first extracts only face regions from the target image and then matches non-face regions in the source image through final steps such as affine transformation and shape refinement, which may cause visual artifacts. Those fake videos are composed of five forgery methods: DeepFakes, Face2Face, FaceSwap, FaceShifter, and Neuraltextures. Each of them contains three quality levels, including uncompressed high-quality (HQ, raw) version, weakly compressed medium-quality (MQ, c23) version, and highly compressed low-quality (LQ, c40) version, respectively. The mechanics behind each DeepFake method are given below.

1) DeepFakes: DeepFakes is an identity manipulation technique by uses an autoencoder to replace the face of the target video with the face of the source video. Moreover, the poisson blending post-processing is utilized to synthesize realistic images further.

2) Face2Face: Face2Face is one kind of facial expression swap method that transfers the expression of a source video to a target video while maintaining the identity of the target person.

3) FaceSwap: FaceSwap belongs to facial identity swap methods. The source person's face is first extracted based on sparsely detected facial features, and a 3D template model is fitted using blend shapes, which are then copied into the target person's facial features, and finally color-corrected for realistic synthetic content.

4) FaceShifter: The task realized by FaceShifter is the face identity transformation in face generation, which means the target face can be replaced with the source face, where the source image provides identity, and the target image provides attributes. The entire face-swapping stage consists of two parts, in the first stage, a GAN-based Adaptive Embedding Integration Network (AEI-Net) is designed for adaptive integration of target attributes. In the second stage, they use the Heuristic Error Acknowledging Network (HEARNet) to handle the facial occlusions and refine the result.

5) Neuraltextures: NeuralTextures proposed a face reenactment method that utilizes rendering methods to modify only the mouth region of the target video while keeping the rest of the region unchanged, making synthetic data more realistic by minimizing adversarial and reconstruction losses.

Celeb-DF. The Celeb-DF dataset contains two released versions (V1 and V2). Compared with V1, Celeb-DF V2 contains more videos and more realistic visual perception. So far, Celeb-DF includes 590 original videos of different ages, races, and genders collected from YouTube, and 5639 corresponding DeepFake videos.

OpenForensics. OpenForensics is a large-scale challenging dataset created for multi-face forgery detection. With rich and explicit facial annotations, this dataset has great research potential in DeepFake prevention and face detection in general. In contrast, this dataset contains various backgrounds and multiple people of different ages, genders, poses, positions, and face occlusions, which is more suitable for real-world scene detection.

Implementation details. We implement all the approaches with PyTorch. First, we perform preprocessing on the dataset. Specifically, for FF++ and Celeb-DF datasets, we use the face detection model MTCNN to locate and crop the faces in the video dataset. In order to preserve more face regions and operation traces, we use conservative cropping to enlarge the face regions by a factor of 1.3 around the tracked face center. Then, the cropped RGB face image is resized to 224×224×3 pixels. At the same time, in order to explore the data differences with different compression levels during detection, we must keep the same position and size when cropping faces under different compression strength versions of the same fake data in FF++. Finally, for each synthetic dataset in FF++ data, we selected 750 fake videos as the training set and the remaining 250 fake videos as the testing set. In order to ensure the balance between real and fake data, we also select 750 real videos as the training set, and the remaining 250 real videos as the testing set. Similarly, for the Celeb-DF v1 dataset, we select 600 videos (300 real, 300 fake) as the training set and 216 videos as the testing set (108 real, 108 fake). For the Celeb-DF v2 dataset, we selected 1400 videos as the training set (700 real, 700 fake) and 380 as the testing set (190 real, 190 fake). In our experiments, all models are trained on frame-level images. Thus, we sample 32 frames for each video, consistent with the literature [3].

Moreover, since the OpenForensics dataset only contains images, we don't need to perform video frame extraction on them. For the preprocessing of this dataset, we first adopted data randomization and then selected 12,000 images as the training set, and 4,000 images as the test set for experiments. Just like the above setting, each image of this dataset was set to 224×224×3.

Considering fairness, we perform the same setting on all models during the training phase. Specifically, for the training protocol, we choose SGD with momentum as the optimizer, and the weight decay is set as 1e-5. In addition, we set the initial learning rate to 0.001, the momentum to 0.9, and the batch size to 8. At the same time, to reduce the burden on computing resources, we train each model for 20 iterations.

Evaluation metrics. For evaluation metrics, following most experimental settings, we use the Accuracy (Acc) and Area Under the Receiver Operating Characteristic Curve (AUC-ROC), commonly used in face forgery detection tasks, to perform verification.

4.2. Qualitative Visual Analysis

In order to more intuitively show that the features extracted by different branches in our proposed method are complementary, we visualize the output results before the fusion of the three branches, as shown in Figure 12. The first column represents the content synthesized by different forgery methods. The second column, Figure 12 (a), shows the local high-frequency information learning process of the Local HiFE branch based on Block-wise DCT. The third column, Figure 12 (b), is the feature heatmap visualization result extracted by the backbone network Xception before fusion. Finally, Figure 12 (c) reveals the global high-frequency information learning process of the Global HiFE branch based on Multi-level DWT. We can find that the features learned by the three branches are different, which is indicated

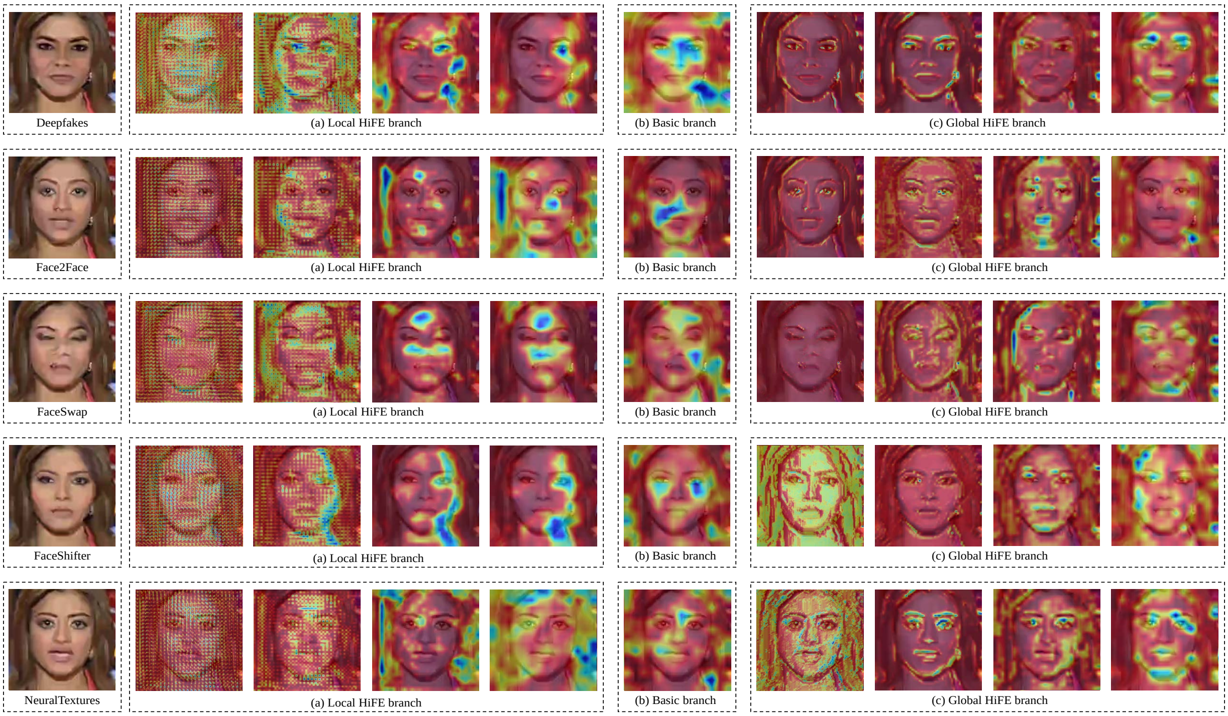


Figure 12: Heatmap visualization of our proposed model in each branch based on CAM [20]. The first column shows the original image (c40 version) of different forgery methods. Figure (a) shows the learning process of the Local High-Frequency Enhancement branch. Figure (b) indicates the feature activation map of the backbone network before fusion. Figure (c) describes the learning process of the Global High-Frequency Enhancement branch.

by the feature heatmap. Therefore, there is a complementary relationship between the three branches, which also lays the foundation for performance improvement through fusion. Next, we separately analyze how the two branches we proposed improve the detection performance.

Impact of Local HiFE branch. Fig. 12 (a) shows the feature learning process of the LHiFE branch based on Block-wise DCT after frequency-aware spatial attention from left to right. We can see that the first picture clearly shows that the importance of different local regions is different. The regions with higher brightness are assigned higher weights, indicating that the model pays more attention to those areas. With the continuous learning process of the network, a connected local region heat map is gradually formed, realizing a local-to-global attention model.

Impact of Global HiFE branch. Fig. 12 (c) illustrates the feature learning process of the GHiFE branch based on Multi-level DWT from left to right. We can observe a progressive enhancement of high-frequency information, transitioning from focusing on fewer local contour details in the first image to ultimately forming broader global details. It also indicates the effectiveness of our multi-level fusion approach in enhancing the extraction of high-frequency detail information.

Based on the visual analysis, we can conclude that our method can effectively extract local and global high-frequency information with complementarity, which means our model can have multi-view capabilities and capture more detailed information. Additionally, from the experimental results in Table 2 and Table 4, it can be observed that our model not only enhances deepfake detection performance at high compression ratios but also shows performance improvement for deepfake content at lower compression ratios. For example, in Table 2, for the uncompressed version (raw) in the FF++ dataset, our algorithm improves by about 0.63% on average; for the lightly compressed version (c23), the average improvement is about 2.99%. In addition, among all SOTA methods in Table 4, the average performance of MAT[74] is the best, and its average performance of all compressed versions in the FF++ dataset is about 94.39%. In comparison, our method HiFE is better, with an average performance of about 94.69%. Even on the OpenForensics

Table 2
Ablation experiments on the performance of LHiFE and GHiFE fusion with backbone Xception and the Two-Stage Cross-Fusion module (the second stage).

Data	Data quality			Acc(%)				Accuracy
Type	raw	c23	c40	Xception [6]	LHiFE(Ours,Xception)	GHiFE(Ours,Xception)	HiFE(Ours,Xception)	Improvement
Deep fakes	✓			99.81	**99.88**	99.66	99.86	0.05↑
		✓		97.41	98.80	99.02	**99.16**	1.75↑
			✓	90.81	93.88	94.24	**94.46**	**3.65↑**
Face2 Face	✓			99.26	99.76	99.49	**99.82**	0.56↑
		✓		96.41	99.13	99.13	**99.45**	3.04↑
			✓	80.94	86.01	85.43	**86.16**	**5.22↑**
Face Swap	✓			98.74	99.83	**99.86**	99.80	1.06↑
		✓		96.62	99.05	98.92	**99.12**	2.50↑
			✓	79.97	87.48	87.28	**87.99**	**8.02↑**
Face Shifter	✓			99.56	99.57	99.56	**99.57**	0.01↑
		✓		96.99	**99.04**	98.98	98.99	2.00↑
			✓	86.01	92.28	90.96	**92.63**	**6.62↑**
Neural Textures	✓			98.33	99.55	99.58	**99.79**	1.46↑
		✓		89.33	94.18	94.59	**94.98**	5.65↑
			✓	65.14	68.82	**69.11**	68.66	**3.52↑**

Table 3
Ablation experiments on the Dual Cross-Attention (DCA) module and the Two-Stage Cross-Fusion module (the first stage). The results (Acc%) show comparisons of the single HiFE branch fused with Xception and after removing the DCA module.

Data	Data quality			Xception	LHiFE	LHiFE	GHiFE	GHiFE
Type	raw	c23	c40		(without DCA)	(Ours,Xception)	(without DCA)	(Ours,Xception)
Deep fakes	✓			99.81	99.68	99.88(0.02↑)	99.66	99.66(0.00↑)
		✓		97.41	98.28	98.80(0.52↑)	98.46	99.02(0.56↑)
			✓	90.81	93.38	93.88(0.50↑)	92.86	94.24(1.38↑)
Face2 Face	✓			99.26	99.08	99.76(0.68↑)	99.47	99.49(0.02↑)
		✓		96.41	98.66	99.13(0.47↑)	98.77	99.13(0.36↑)
			✓	80.94	84.31	86.01(1.70↑)	84.36	85.43(1.07↑)
Face Swap	✓			98.74	99.76	99.83(0.07↑)	99.63	99.86(0.23↑)
		✓		96.26	98.38	99.05(0.67↑)	98.41	98.92(0.51↑)
			✓	79.97	86.29	87.48(1.19↑)	85.71	87.28(1.57↑)
Face Shifter	✓			99.56	99.55	99.57(0.02↑)	99.56	99.56(0.00↑)
		✓		96.99	98.57	99.04(0.47↑)	98.35	98.98(0.63↑)
			✓	86.01	90.46	92.28(1.82↑)	89.38	90.96(1.58↑)
Neural Textures	✓			98.33	99.06	99.55(0.49↑)	99.29	99.58(0.29↑)
		✓		89.33	93.78	94.18(0.40↑)	93.61	94.59(0.98↑)
			✓	65.14	67.71	68.82(1.11↑)	67.45	69.11(1.66↑)

dataset, which applies various perturbations to better simulate natural scenes, our method outperforms existing state-of-the-art methods (Table 7), which also proves that our method can alleviate the indistinguishability problem caused by compression during transmission to a certain extent.

4.3. Quantitative Ablation Analysis

To quantitatively illustrate the effectiveness of our proposed method, we conduct ablation study experiments to explore the performance of different branches as well as fusion modules.

As shown in Table 2, we compare the performance of the original backbone network Xception, the GHiFE branch embedded with the backbone, the LHiFE branch embedded with the backbone, and the final network HiFE. The best detection results are highlighted in bold. Apparently, the results in the fourth and fifth columns of Table 2 confirm the effectiveness of merging information from individual branches, leading to a significant improvement in the performance of the backbone network. Moreover, in comparison, the network with the fusion of the three branches performs better on low-quality data (c40), with an average performance improvement of approximately 5.41%. This also indicates that

Table 4
Comparison of the detection performance (Acc%) of our method and SOTA methods based on three different compression strengths in the FF++ dataset.

Data Type	Data quality raw	c23	c40	F3Net (FAD+LFS)	CEVIT	Two stream	CVIT	EfficientNet AutoAttB4	EfficientNet AutoAttB4ST	Meso Net	Meso Inception4	MAT	MCX-API	HiFE (Ours,Xception)
Deep fakes	✓			**99.98**	99.61	99.75	99.74	99.88	92.45	95.81	96.23	99.50	99.31	99.86
		✓		98.49	97.69	98.16	96.65	**99.18**	80.01	81.18	91.36	98.39	94.89	99.16
			✓	93.26	90.45	92.29	90.38	93.40	69.38	73.36	84.23	94.41	87.68	**94.46**
Face2 Face	✓			99.73	99.76	99.64	99.84	**99.90**	83.01	62.93	97.21	99.48	99.35	99.82
		✓		98.36	98.27	98.04	97.69	99.08	72.59	81.84	87.61	98.94	96.73	**99.45**
			✓	84.29	82.43	83.78	84.79	84.34	60.15	72.55	69.83	86.00	81.07	**86.16**
Face Swap	✓			99.34	99.17	99.72	98.66	99.79	84.15	72.15	95.34	99.64	97.47	**99.80**
		✓		97.65	97.79	97.17	97.18	98.88	77.25	71.48	86.13	98.49	94.75	**99.12**
			✓	84.43	80.89	83.72	83.18	86.99	60.18	62.03	59.24	87.69	81.20	**87.99**
Face Shifter	✓			99.57	99.45	99.54	99.56	**99.58**	89.49	90.97	96.38	99.56	99.41	99.57
		✓		98.23	97.18	97.28	96.29	98.43	76.20	85.56	95.23	98.56	93.35	**98.90**
			✓	88.02	88.88	88.08	86.77	88.83	66.66	69.16	80.16	91.44	86.21	**92.63**
Neural Textures	✓			99.43	98.94	**99.95**	99.15	99.85	82.07	55.26	91.78	99.66	98.28	99.79
		✓		92.19	92.13	90.00	89.55	94.58	70.24	62.09	66.26	94.26	84.36	**94.98**
			✓	67.92	68.14	69.36	66.91	67.72	57.64	61.59	62.23	**69.84**	64.00	68.66

the proposed Two-Stage Cross-Fusion module (the second stage) can better integrate information from three domains. Additionally, the advantages of our proposed method are further evident in the case of low compression strength data (c23), with an average accuracy improvement of about 2.99%. It also enhances the accuracy of uncompressed original data (raw) by approximately 0.63%. These experimental results validate the effectiveness of our two branches and fusion module on DeepFake data, particularly on low-quality data.

In order to verify whether the Dual Cross-Attention module and our Two-Stage Cross-Fusion module (the first stage) effectively fuse information from two different domains, we performed ablation experiments, as shown in Table 3. We removed the Dual Cross-Attention module and only retained the backbone, single HiFE branch, and the Two-Stage Cross-Fusion module (the first stage) for experimental verification. Compared with the backbone network Xception in the third column of Table 3, the results in the fourth (LHiFE (without DCA)) and sixth (GHiFE (without DCA)) columns reflect that the Two-Stage Cross-Fusion module (the first stage) we designed has greatly improved the detection performance. In addition, the results in the fifth and seventh columns prove that the embedding of Dual Cross-Attention can further improve the discriminative ability of the algorithm.

4.4. Comparative Experiments

In this section, we compare the proposed method with the state-of-the-art method to verify the effectiveness of our proposed method from three perspectives.

Different Compressive Strengths. Considering that the original intention of our method is to improve the detection performance of compressed DeepFake data, we first analyze the experimental results of our proposed method and the state-of-the-art at different compression strengths. As shown in Table 4, we conduct experiments on five forgy methods which from FF++ data. Each fake method includes three compression levels: raw (HQ), c23 (MQ), and c40 (LQ), respectively. We selected ten SOTA methods to compare with our method. Among them, F3Net [51] distinguishes real from fake by constructing dual-stream frequency-aware clues. The author uses DCT as the applied frequency-domain transform. One stream manually divides the frequency sub-bands and constructs a frequency-aware decomposition (FAD) image. The other stream uses local frequency statistical (LFS) features to achieve classification. Coccomini *et al.* [8] combined Vision Transformers with convolutional EfficientNet (CEVIT), where EfficientNet is used as a feature extractor, and the obtained features are sent to Transformer for classification. Luo *et al.* [42] designed a two-stream architecture (Two stream), where the residual-guided spatial features and multi-scale high-frequency features are fused for final classification. Wodajo and Atnafu [63] proposed a Convolutional Vision Transformer (CVIT) for the detection of DeepFakes, which combines a Convolutional Neural Network (CNN) and Vision Transformer (ViT). Among them, CNN extracts learnable features, while ViT takes the learned features as input and classifies them using an attention mechanism. In addition, both EfficientNetAutoAttB4 and EfficientNetAutoAttB4ST [3] are improved versions based on EfficientNetB4. The former integrates the attention mechanism into the network, and the latter uses the Siamese Training mode. Afchar *et al.* [1] proposed to use a deep neural network with fewer layers for deepfake detection, and constructed two structures: MesoNet and MesoInception4. They argue that image noise-based microscopic analysis cannot be applied in compressed video environments where image noise is severely degraded. Moreover, at a higher semantic level, it is difficult for the human eye to distinguish fake images. Therefore, they choose

Table 5
Comparison of time complexity (FLOPs on one sample) and the number of model parameters (Params) of our proposed method and the best SOTA method (MAT).

Model	Params	FLOPs
MAT	417.63M	76.41G
LHiFE(Ours,Xception)	27.46M	7.32G
GHiFE(Ours,Xception)	24.03M	5.24G
HiFE(Ours,Xception)	47.55M	11.29G

Table 6
Comparison of the model size between our proposed method and other SOTA methods.

Model	Size
Xception	159MB
F3Net	321MB
Two stream	406MB
CVIT	679MB
EfficientNetAutoAttB4	134MB
EfficientNetAutoAttB4ST	67.7MB
MesoNet	198KB
MesoInception4	198KB
CEVIT	749MB
MAT	3.11GB
MCX-API	183MB
LHiFE(Ours,Xception)	219MB
GHiFE(Ours,Xception)	193MB
HiFE(Ours,Xception)	363MB

mid-level semantics for recognition. Zhao *et al.* [74] formulated deepfake detection as a fine-grained classification problem and proposed a multi-attention deepfake detection network (MAT). By constructing multiple spatial attention heads, texture feature enhancement block, and feature aggregation module, they effectively improved the detection performance of low-quality data. Last but not least, Xu *et al.* [67] proposed Multi-Channel Xception Attention Pairwise Interaction (MCX-API), which utilizes paired images for learning in a fine-grained manner and constructs multiple color channels. Compared with the above methods, through local and global high-frequency information enhancement, our method automatically learns and enhances the residual high-frequency information on low-quality images. The experimental results in Table 4 show that our method achieved better performance in most cases under various compression versions of FF++, especially on low-quality data (c40). Furthermore, on the other hand, we utilize the AUC-ROC curve as another measure of classifier performance. As shown in Figure 13, we plot the AUC-ROC curves of the remaining comparison methods and our proposed method on the FF++ dataset, where the classifier with a larger AUC score has better performance. Similarly, we obtained the experimental results of all detection methods under different compression strengths. From the results in the figure, we can see that our method outperforms other methods. Specifically, the AUC of our proposed method is larger, and the ROC curve is sharper, which means that our approach has better classification performance. Especially on highly compressed data (c40), our method significantly outperforms other methods.

Furthermore, it is worth emphasizing that our model is lightweight. As depicted in Table 5, compared with the best-performing state-of-the-art method (MAT), our model has two significant advantages: lower FLOPs and fewer model parameters, which indicates faster inference, simpler, and easier to train. In addition, we provide the comparison results of model sizes of all methods in Table 6 to illustrate that our method is more convenient to deploy in resource-constrained environments.

Different Dataset. To verify that our proposed method is not only effective for the FF++ dataset, in this part, we conduct additional experiments on other datasets. The experimental results are shown in Table 7. Considering that we aim to improve the accuracy of low-quality fake data, as a further supplementary illustration, we conduct experiments on three different compressed versions of the entire FF++ dataset, including five fake patterns. Furthermore, we select three more challenging datasets to illustrate the practical generalizability of our method. Experiments were carried

Table 7
Performance comparison of detection methods based on FF++, Celeb-DF and OpenForensics.

Method	Dataset					
	FF++ (raw)	FF++ (c23)	FF++ (c40)	Celeb-DF V1	Celeb-DF V2	OpenForensics
Xception [6]	98.88	90.46	69.97	94.76	95.30	98.70
F3Net(FAD) [51]	98.69	89.22	69.68	94.40	94.12	98.93
F3Net(LFS) [51]	90.03	85.17	62.52	80.61	80.48	93.90
F3Net(FAD+LFS) [51]	99.10	89.56	70.79	94.46	94.15	98.38
Two stream [42]	**99.66**	78.96	66.13	85.26	88.87	95.73
CVIT [63]	91.09	74.29	65.72	81.52	90.86	96.53
EfficientNetAutoAttB4 [3]	99.58	91.96	70.02	95.17	95.33	98.23
EfficientNetAutoAttB4ST [3]	69.66	61.41	55.06	76.24	69.39	86.10
MesoNet [1]	66.13	62.34	50.31	72.44	71.73	89.53
MesoInception4 [1]	75.24	68.87	65.86	78.94	82.83	92.48
CEVIT [8]	99.46	91.14	69.29	92.62	94.40	96.93
MAT [74]	98.63	90.34	69.84	95.23	95.63	98.53
MCX-API [67]	87.67	67.49	62.85	82.71	87.73	95.48
LHiFE (Ours,Xception)	98.97	92.82	71.51	**96.33**	95.56	98.35
GHiFE (Ours,Xception)	99.31	92.31	71.05	95.01	95.23	98.75
HiFE (Ours,Xception)	99.36	**92.83**	**71.84**	95.69	**96.64**	**99.03**

Table 8
Experimental results of the proposed method based on different backbone networks.

Model	Dataset						Avg.
	FF++ (raw)	FF++ (c23)	FF++ (c40)	Celeb-DF V1	Celeb-DF V2	OpenForensics	
Xception	98.88	90.46	69.97	94.76	95.30	98.70	91.35
HiFE(Ours,Xception)	99.36(**0.48**↑)	92.83(**2.37**↑)	71.84(**1.87**↑)	95.69(**0.93**↑)	96.64(**1.34**↑)	99.03(**0.33**↑)	92.57(**1.22**↑)
EfficientNetB4	99.29	91.23	69.39	94.84	95.58	97.38	91.29
HiFE(Ours,EfficientNetB4)	99.61(**0.32**↑)	91.96(**0.73**↑)	72.54(**3.15**↑)	97.12(**2.28**↑)	96.83(**1.25**↑)	98.60(**1.22**↑)	92.78(**1.49**↑)

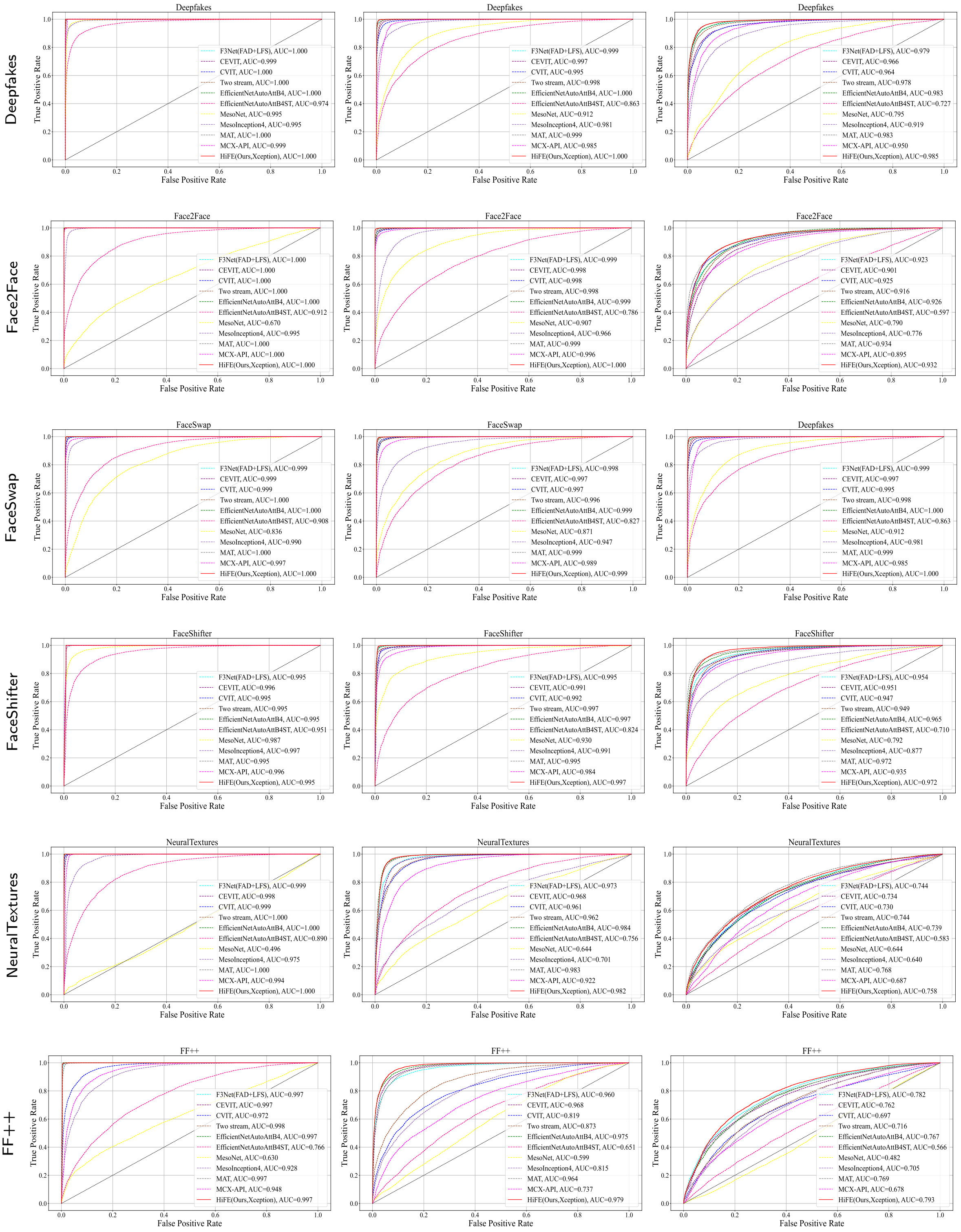
out on two versions of Celeb-DF, as well as the OpenForensics dataset. Similarly, we mainly select ten methods for comparison (except for Xception), and perform ablation experiments. From Table 7, it can be concluded that our method works best on the whole FF++ dataset with low quality (c40). Moreover, it can be seen from Table 7 that our proposed method outperforms most methods even on the highly realistic Celeb-DF dataset and OpenForensics dataset, which is in the wild scenes. Same as the experiments in the previous section, we plot the AUC-ROC curves of all models on these datasets, as shown in Figure 14. The superiority of our method is very obvious, which illustrates that in the face of more challenging fake data, our method also has advantages compared with other SOTA methods.

Different Backbone Networks. To illustrate the flexibility and transferability of our model, except Xception, we further conduct experiments using EfficientNetB4 as the backbone network, as shown in Table 8. Considering that our purpose is to improve the accuracy of low-quality fake data, we still conduct experimental verification on three different compression levels of aligned data in FF++. Here, we select the entire FF++ dataset for experimental illustration. In addition, we also select the Celeb-DF and OpenForensics datasets for further experimental supplements to prove that our method is not only applicable to FF++. Table 8 shows that our method can perfectly adapt to EfficientNetB4. Even though EfficientNetB4 has achieved extremely excellent performance, we can still further improve the performance with it as the backbone network, which further verifies the effectiveness and flexibility of our proposed method.

5. Conclusion

This paper focuses on the detection problem of low-quality DeepFake data. To address this issue, we conduct an empirical study between low-quality faces with high compression ratios and corresponding high-quality faces; then we draw two essential observations: first, the high-frequency information of low-quality data is different from high-quality data; second, we design information exchange strategies between low-quality data and high-quality data, respectively, and found that after high-frequency information exchange, new low-quality data were easier to identify.

Therefore, we propose a novel high-frequency enhancement algorithm, which we called HiFE, integrating the high-frequency feature enhancement step into the network architecture as follows: (1) a global high-frequency enhancement



(a)raw (b)c23 (c)c40

Figure 13: Comparison of AUC results of various detection methods on FF++ dataset with different compress ratio.

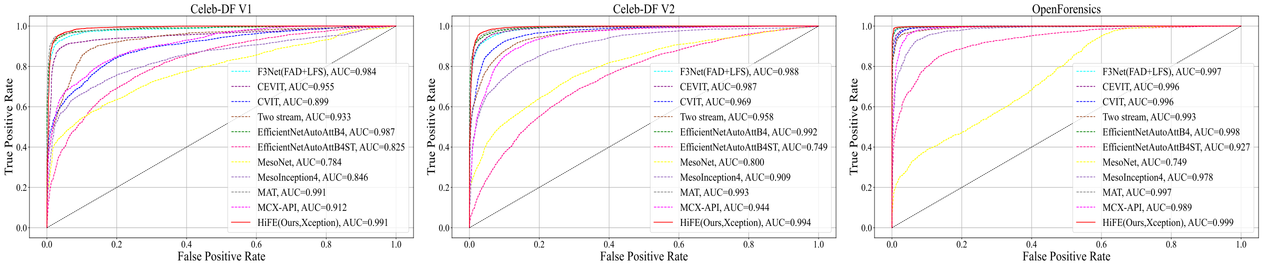


Figure 14: Comparison of AUC results of various detection methods on Celeb-DF (V1 and V2) and OpenForensics datasets.

branch is used to enhance the global high-frequency information by constructing Multi-level DWT and cascade-residual-based multi-level fusion strategies; (2) a local high-frequency enhancement branch based on Block-wise DCT by constructing a frequency-aware multi-spatial attention strategy reflects the importance of its frequency-domain information in different local spatial domains. Through the proposed fusion strategy, we combine the global high-frequency enhancement branch and the local high-frequency enhancement branch with the backbone network to obtain our HiFE system. Extensive comparative and ablation experiments show that our proposed framework effectively improved the performance of DeepFake detection, especially for low-quality data with high compression rates. Our method achieves state-of-the-art on the FF++, Celeb-DF, and OpenForensics datasets.

Although our approach focuses on investigating the reasons for performance degradation in low-quality content and has significantly improved performance, we have not considered research on the generalization ability in the case of cross-dataset, which is also regarded as one of the challenges in DeepFake detection. Furthermore, we only utilize spatial information and have not explored temporal inconsistencies between frames. Therefore, we will address this issue in future work by examining it from both temporal and spatial perspectives.

Acknowledgments This work is partially supported by the China Scholarship Council (CSC, No.202206290093), by SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU, and by the Key Research and Development Program of Shaanxi (Nos. 2023ZDLGY16, 2021ZDLGY15-01, 2021ZDLGY09-04, and 2021GY-004). All authors contributed to this paper. All authors read and approved the final manuscript. The contribution list is below.
Jie Gao: Investigation, Data Prepare, Visualization, Methodology, Software, Writing – original draft preparation.
Zhaoqiang Xia: Supervision, Writing – reviewing & editing.
Gian Luca Marcialis: Supervision, Writing – reviewing & editing.
Chen Dang: Experimental Verification, Writing – reviewing & correcting.
Jing Dai: Resources, Writing – reviewing.
Xiaoyi Feng: Resources, Supervision, Writing – reviewing.

Declarations The authors declare that there is no conflict of interest in this manuscript.

Data availability The data that support the findings of this study are available here:

1. FaceForensics++
   `https://github.com/ondyari/FaceForensics`
2. Celeb-DF
   `https://github.com/yuezunli/celeb-deepfakeforensics`
3. OpenForensics
   `https://sites.google.com/view/ltnghia/research/openforensics`

References

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *Proceedings of the International Workshop on Information Forensics and Security (WIFS)*, pages 1–7. IEEE, 2018.

[2] Nicolas Beuve, Wassim Hamidouche, and Olivier Déforges. Waterlo: Protect images from deepfakes using localized semi-fragile watermark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 393–402, 2023.

[3] Nicolo Bonettini, Edoardo Daniele Cannas, Sara Mandelli, Luca Bondi, Paolo Bestagini, and Stefano Tubaro. Video face manipulation detection through ensemble of cnns. In *Proceedings of the international conference on pattern recognition (ICPR)*, pages 5012–5019. IEEE, 2021.

[4] Han Chen, Yuezun Li, Dongdong Lin, Bin Li, and Junqiang Wu. Watching the big artifacts: Exposing deepfake videos via bi-granularity artifacts. *Pattern Recognition*, 135:109179, 2023.

[5] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1081–1088, 2021.

[6] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1251–1258. IEEE, 2017.

[7] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[8] Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. Combining efficientnet and vision transformers for video deepfake detection. In *Proceedings of the International Conference on Image Analysis and Processing (ICIAP)*, pages 219–229. Springer, 2022.

[9] Sara Concas, Simone Maurizio La Cava, Giulia Orrù, Carlo Cuccu, Jie Gao, Xiaoyi Feng, Gian Luca Marcialis, and Fabio Roli. Analysis of score-level fusion rules for deepfake detection. *Applied Sciences*, 12(15):7365, 2022.

[10] Sara Concas, Gianpaolo Perelli, Gian Luca Marcialis, and Giovanni Puglisi. Tensor-based deepfake detection in scaled and compressed images. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3121–3125. IEEE, 2022.

[11] Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3994–4004, 2023.

[12] Ricard Durall, Margret Keuper, Franz-Josef Pfreundt, and Janis Keuper. Unmasking deepfakes with simple features, 2019.

[13] Sheldon Fung, Xuequan Lu, Chao Zhang, and Chang-Tsun Li. Deepfakeucl: Deepfake detection via unsupervised contrastive learning. In *2021 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2021.

[14] Jie Gao, Sara Concas, Giulia Orrù, Xiaoyi Feng, Gian Luca Marcialis, and Fabio Roli. Generalized deepfake detection algorithm based on inconsistency between inner and outer faces. In *International Conference on Image Analysis and Processing*, pages 343–355. Springer, 2023.

[15] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, pages 2672–2680, 2014.

[16] Ijaz Ul Haq, Khalid Mahmood Malik, and Khan Muhammad. Multimodal neurosymbolic approach for explainable deepfake detection. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023.

[17] Javier Hernandez-Ortega, Ruben Tolosana, Julian Fierrez, and Aythami Morales. Deepfakeson-phys: Deepfakes detection based on heart rate estimation. *arXiv preprint arXiv:2010.00400*, 2020.

[18] Juan Hu, Xin Liao, Wei Wang, and Zheng Qin. Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1089–1102, 2022. doi: 10.1109/TCSVT.2021.3074259.

[19] Sahar Husseini and Jean-Luc Dugelay. A comprehensive framework for evaluating deepfake generators: Dataset, metrics performance, and comparative analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 372–381, 2023.

[20] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021.

[21] Minguk Kang, Joonghyuk Shin, and Jaesik Park. Studiogan: a taxonomy and benchmark of gans for image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *Proceedings of International Conference on Learning Representations (ICLR)*, pages 1–14, 2014.

[24] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018.

[25] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 3677–3685, 2017.

[26] Nicolas Larue, Ngoc-Son Vu, Vitomir Struc, Peter Peer, and Vassilis Christophides. Seeable: Soft discrepancies and bounded contrastive learning for exposing deepfakes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21011–21021, 2023.

[27] T.-N. Le, H.H. Nguyen, J. Yamagishi, and I. Echizen. Openforensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild. In *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV2021)*, page 10117–10127. IEEE, 2021.

[28] Gen Li, Xianfeng Zhao, and Yun Cao. Forensic symmetry for deepfakes. *IEEE Transactions on Information Forensics and Security*, 18: 1095–1110, 2023.

[29] H Li, B Li, S Tan, and J Huang. Identification of deep network generated images using disparities in color components. *arXiv preprint arXiv:1808.07276*, pages 1–26, 2018.

[30] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, pages 1–11, 2019.

[31] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5001–5010. IEEE, 2020.

[32] Y Li, X Yang, P Sun, H Qi, and S Celeb-DF Lyu. A large-scale challenging dataset for deepfake forensics. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA*, pages 14–19. IEEE, 2020.

[33] Ying Li, Shan Bian, Chuntao Wang, Kemal Polat, Adi Alhudhaif, and Fayadh Alenezi. Exposing low-quality deepfake videos of social network service using spatial restored detection framework. *Expert Systems with Applications*, page 120646, 2023.

[34] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, pages 1–7, 2018.

[35] Xin Liao, Yumei Wang, Tianyi Wang, Juan Hu, and Xiaoshuai Wu. Famm: Facial muscle motions for detecting compressed deepfake videos over social networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[36] Baoping Liu, Bo Liu, Ming Ding, Tianqing Zhu, and Xin Yu. Ti2net: Temporal identity inconsistency network for deepfake detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4691–4700, 2023.

[37] Chi Liu, Huajie Chen, Tianqing Zhu, Jun Zhang, and Wanlei Zhou. Making deepfakes more spurious: Evading deep face forgery detection via trace removal attack. *IEEE Transactions on Dependable and Secure Computing*, 20(6):5182–5196, 2023. doi: 10.1109/TDSC.2023.3241604.

[38] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 772–781, 2021.

[39] Kai Liu, Bicheng Li, and Jiale Li. Deep face-swap model combining attention mechanism and cyclegan. In *Journal of Physics: Conference Series*, volume 2278, page 012037. IOP Publishing, 2022.

[40] Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8060–8069. IEEE, 2020.

[41] Tao Luo, Zheng Ma, Zhi-Qin John Xu, and Yaoyu Zhang. Theory of the frequency principle for general deep neural networks. *CoRR*, abs/1906.09235, 2019. URL `http://arxiv.org/abs/1906.09235`.

[42] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16317–16326. IEEE, 2021.

[43] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *Proceedings of the Winter Applications of Computer Vision Workshops (WACVW)*, pages 83–92. IEEE, 2019.

[44] Nesryne Mejri, Enjie Ghorbel, and Djamila Aouada. Untag: Learning generic features for unsupervised type-agnostic deepfake detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[45] Changtao Miao, Zichang Tan, Qi Chu, Huan Liu, Honggang Hu, and Nenghai Yu. F 2 trans: High-frequency fine-grained transformer for face forgery detection. *IEEE Transactions on Information Forensics and Security*, 18:1039–1051, 2023.

[46] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)*, 54(1):1–41, 2021.

[47] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *Proceedings of the 28th ACM International Conference on Multimedia(ACMMM)*, pages 2823–2832, 2020.

[48] Karima Omar, Rasha H Sakr, and Mohammed F Alrahmawy. An ensemble of cnns with self-attention mechanism for deepfake video detection. *Neural Computing and Applications*, 36(6):2749–2765, 2024.

[49] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 484–492, 2020.

[50] Hua Qi, Qing Guo, Felix Juefei-Xu, Xiaofei Xie, Lei Ma, Wei Feng, Yang Liu, and Jianjun Zhao. Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms. In *Proceedings of the 28th ACM International Conference on Multimedia (ACMMM)*, pages 4318–4327, 2020.

[51] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 86–103. Springer, 2020.

[52] Tong Qiao, Shichuang Xie, Yanli Chen, Florent Retraint, and Xiangyang Luo. Fully unsupervised deepfake video detection via enhanced contrastive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[53] Xuejian Rong, Denis Demandolx, Kevin Matzen, Priyam Chatterjee, and Yingli Tian. Burst denoising via temporally shifted wavelet transforms. In *Proceedings of the European Conference on Computer Vision(ECCV)*, pages 240–256. Springer, 2020.

[54] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 1–11. IEEE, 2019.

[55] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18720–18729. IEEE, 2022.

[56] Simranjeet Singh, Rajneesh Sharma, and Alan F Smeaton. Using gans to synthesise minimum training data for deepfake generation. *arXiv preprint arXiv:2011.05421*, 2020.

[57] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16515–16525, 2022.

[58] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2387–2395. IEEE, 2016.

[59] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

[60] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion (IF)*, 64:131–148, 2020.

[61] R Wang, L Ma, F Juefei-Xu, X Xie, J Wang, and Y Liu. Fakespotter: A simple baseline for spotting ai-synthesized fake faces. *arXiv preprint arXiv:1909.06122*, pages 1–8, 2019.

[62] Yaohui Wang and Antitza Dantcheva. A video is worth more than 1000 lies. comparing 3dcnn approaches for detecting deepfakes. In *2020 15Th IEEE international conference on automatic face and gesture recognition (FG 2020)*, pages 515–519. IEEE, 2020.

[63] Deressa Wodajo and Solomon Atnafu. Deepfake video detection using convolutional vision transformer. *CoRR*, abs/2102.11126, 2021. URL `https://arxiv.org/abs/2102.11126`.

[64] Simon Woo et al. Add: Frequency attention and multi-view based knowledge distillation to detect low-quality compressed deepfake images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 122–130, 2022.

[65] Jiahui Wu, Yu Zhu, Xiaoben Jiang, Yatong Liu, and Jiajun Lin. Local attention and long-distance interaction of rppg for deepfake detection. *The Visual Computer*, 40(2):1083–1094, 2024.

[66] Mingqing Xiao, Shuxin Zheng, Chang Liu, Yaolong Wang, Di He, Guolin Ke, Jiang Bian, Zhouchen Lin, and Tie-Yan Liu. Invertible image rescaling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 126–144. Springer, 2020.

[67] Ying Xu, Kiran Raja, Luisa Verdoliva, and Marius Pedersen. Learning pairwise interaction for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 672–682, 2023.

[68] Zhi-Qin John Xu. Frequency principle: Fourier analysis sheds light on deep neural networks. *Communications in Computational Physics(COMMUN COMPUT PHYS)*, 28(5):1746–1767, 2020.

[69] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019.

[70] Peipeng Yu, Zhihua Xia, Jianwei Fei, and Yujiang Lu. A survey on deepfake video detection. *Iet Biometrics*, 10(6):607–624, 2021.

[71] Yang Yu, Xiaohui Zhao, Rongrong Ni, Siyuan Yang, Yao Zhao, and Alex C Kot. Augmented multi-scale spatiotemporal inconsistency magnifier for generalized deepfake detection. *IEEE Transactions on Multimedia*, (99):1–13, 2023.

[72] Li Zhang, Dezong Zhao, Chee Peng Lim, Houshyar Asadi, Haoqian Huang, Yonghong Yu, and Rong Gao. Video deepfake classification using particle swarm optimization-based evolving ensemble models. *Knowledge-Based Systems*, page 111461, 2024.

[73] Yaoyu Zhang, Zhi-Qin John Xu, Tao Luo, and Zheng Ma. Explicitizing an implicit bias of the frequency principle in two-layer neural networks. *CoRR*, abs/1905.10264, 2019. URL `http://arxiv.org/abs/1905.10264`.

[74] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2185–2194. IEEE, June 2021.

[75] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Two-stream neural networks for tampered face detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1831–1839. IEEE, 2017.