Face Forgery Detection Based on Fine-grained Clues and Noise Inconsistency

Dengyong Zhang, Ruiyi He, Xin Liao, Feng Li, Jiaxin Chen, and Gaobo Yang

***Abstract*—Deepfake detection has gained increasingly research attentions in media forensics, and a variety of works have been produced. However, subtle artifacts might be eliminated by compression, and the Convolutional Neural Networks (CNN) based detectors are invalidated for fake face images with compression. In this work, we propose a two-stream network for deepfake detection. We observed that high-frequency noise features and spatial features are inherently complementary to each other. Thus, both spatial features and high-frequency noise features are exploited for face forgery detection. Specifically, we design a Double-Frequency Transformer Module (DFTM) to guide the learning of spatial features from local artifact regions. To effectively fuse spatial features and high-frequency noise features, a Dual Domain Attention Fusion Module (DDAFM) is designed. We also introduce a local relationship constraint loss, which requires only image-level labels, for model training. We evaluate the proposed approach on five large-scale benchmark datasets, and extensive experimental results demonstrate the proposed approach outperforms most SOTA works. Code will be provided at https://github.com/hryyyy/HILIF.**

***Impact Statement*—In recent years, there has been a noticeable trend of an increasing number of individuals turning to social media as their primary source of information. Due to the rapid advancements in deep learning, manipulating faces has become easily achievable, allowing individuals to manipulate and disseminate altered images in the media. The malicious manipulation of faces poses significant threats to personal privacy, social security, and national security. Given that most mainstream media utilize implicit compression, the compression process tends to dilute traces of forgery, making detection more challenging. This poses a significant obstacle to effectively identifying manipulated content within the compressed domain. The techniques introduced in this study can effectively identify low-quality fake content within the compressed domain. Simultaneously, generalization capabilities are essential for forgery detection in real scenarios. Our model demonstrates good generalization performance, making it more suitable for detecting forgeries in real-world applications. In contrast to alternative methods, our deep learning model is characterized by its lightweight nature, making it well-suited for practical use in real-life scenarios.**

***Index Terms*—Face Forgery Detection, Frequency Domain, Dual Domain Attention Fusion, Transformer, Constraint Loss**

I. Introduction

In recent years, artificial intelligence (AI) has made continuous developments. Especially, various generative models have enabled Deepfake to generate scarily-real face images. These photo-realistic face images can be used in some applications such as entertainment and film production. Nevertheless, the deepfake-enabled fake face images can also be used for malicious and illegal purposes such as spreading misinformation and misleading public opinions. Therefore, there is a strong imperative to develop deepfake detection, which exposes AI-powered face forgeries, in the field of multimedia information security.

Most early deepfake detection works [1]–[3] directly learned spatial features to expose the artifacts in fake faces, while some works learned features from residual domain [4], [5] to enhance detection accuracy and generalization capability. These spatial-domain and residual-domain features expose prominent artifacts and inherent inconsistencies in face forgeries. Image noises, which can be considered as an intrinsic pattern specific to hardware devices, can serve as the clue for image forgery detection. Moreover, noise is usually extracted through residual operations. These spatial and residual domain based works can achieve desirable detection results on high-quality datasets, but their performances degrade seriously when dealing with low-quality images. To address the issue, some works [6]–[9] have been presented to exploit frequency-domain features. Some works [6], [9] supposed that the artifacts in fake images are alleviated by some operation such as JPEG compression, making them be imperceptible in RGB domain but still detectable in frequency domain. While the frequency-domain based works achieve substantial performance gains for low-quality images, they are often accompanied by a degraded generalization.

To address the afore-mentioned challenges, we investigate the interplay between different feature domains, with a particular focus on the complementary nature of both spatial and noise domains. The spatial domain features are discriminative for deepfake detection, while the noises based features provide auxiliary information from the inconsistency perspective. For deepfake detection in complex scenarios, fusing features from these different domains proves to be an effective strategy for enhancing detection accuracy and robustness. Moreover, to address the issue associated with low-quality compressed images, we leverage the frequency domain to guide the spatial

This project was supported by the National Natural Science Foundation of China (Grant No. 62172059, 62402062, U22A2030) ; National Key RD Program of China (Grant No. 2022YFB3103500), Hunan Provincial Funds for Distinguished Young Scholars (Grant No. 2024JJ2025), Hunan Provincial Key RD Program (Grant No. 2024AQ2027) and Natural Science Foundation of Changsha City, China(Grant No. kq2402031).(Corresponding author: Xin Liao)

Dengyong Zhang, Ruiyi He, Feng Li, Jiaxin Chen are with Hunan Provincial Key Laboratory of Intelligent Processing of Big Data on Transportation, Changsha University of Science and Technology, Changsha, 410114, China; School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha, 410114, China (e-mail: zhdy@csust.edu.cn, hry@stu.csust.edu.cn, lif@csust.edu.cn, jxchen@csust.edu.cn).

Xin Liao and Gaobo Yang are with Hunan University, Changsha 410082, China (e-mail: xinliao@hnu.edu.cn; yanggaobo@hnu.edu.cn)

This article has been accepted for publication in IEEE Transactions on Artificial Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAI.2024.3455311

domain, capturing subtle artifacts in fake faces under compressed conditions. We improve the self-attention mechanism in Transformer [10] to facilitate the guidance from the frequency domain to the spatial domain. That is, we propose a two-stream network based on the spatial domain and the noise domain, using the frequency domain to guide the spatial domain to capture finer artifacts.

Many existing works regard deepfake detection as a binary classification problem, employing the softmax loss function to supervise CNN training. However, a single softmax Loss may be insufficient for distinguishing various forgeries. Recent studies [11], [12] have observed the softmax Loss function's limitations and proposed alternative loss functions for refining network training. Thus, we enhance the local relationship-constrained loss function introduced by Li et al. [12], optimizing our network training by considering the diversity of face forgeries.

In this work, we present a deepfake detection framework that leverages spatial information guided by frequency domain clues and incorporates noise information. The framework comprises spatial and noise flows. Double-Frequency Transformer Module (DFTM), which exploits high-frequency signals to guide the learning of spatial information from local artifacts, is employed for the spatial flow. In contrast, the SRM, a high-pass filter [13] used to extract noise flow information. Furthermore, a Dual Domain Attention Fusion Module (DDAFM) is utilized to facilitate the interactive combination of spatial and noise information. We also optimize the local relationship-constrained and cross-entropy loss functions for network training. This study's main contributions are explained as follows:

- We introduce DFTM, which leverages frequency-domain information to guide spatial feature learning from detailed artifacts.
- We utilize DDAFM to completely fully exploit the complementarity between spatial and noise features.
- The local relationship constraint loss function is enhanced by computing block-based similarity and amplifying the differences between artifact and normal areas, enabling the model to learn more from discriminative regions.
- We conduct comprehensive evaluations on several benchmark datasets. The extensive experimental results and visualizations show that the proposed approach outperforms state-of-the-art methods on several widely-used datasets.

Furthermore, our work addresses some of the previous studies' limitations:

- Our method employs sliding block Discrete Cosine Transform (DCT) filtering to thoroughly and meticulously explore fine-grained frequency domain forgery clues.
- For fine-grained high-frequency forgery clues, our method is highly adaptable for identifying fine-grained high-frequency forgery clues and is capable of fully exploring correlations between different regions.
- Fine-grained high-frequency forgery clues and RGB features are comprehensively modeled and integrated at different scales. This approach enhances the representation of high-frequency fine-grained forgery clues and captures the correlation between high-frequency and RGB information.

The remainder of this work is organized as follows: Section II introduces related work. Section III describes the feasibility analysis of our approach. Section IV presents the proposed approach. Section V provides the experimental results. Section VI provides a visual analysis. Section VII concludes this work.

II. Related Work

A. Spatial-Based Deepfake Detection

To counter the proliferation of face forgeries, numerous detection methods have been proposed. Zhao et al. [14] formulated deepfake detection as a fine-grained classification problem and proposed a new multi-attentional deepfake detection network, which aggregates the low-level textural feature and high-level semantic features guided by the attention maps. Dang et al. [15] presented to exploit the learned attention maps to highlight the informative regions, further improving the binary classification (genuine face v. fake face). Guo et al. [16] presented a lightweight dynamic fusion network, namely LDFnet, for deepfake detection by integrating local artifacts and global texture information.

B. Frequency-Based Deepfake Detection

Given the effectiveness of frequency information in deepfake detection, several studies have explored the complementary nature of spatial-domain and frequency-domain features. For example, Qian et al. [6] introduced the F3-Net for face forgery detection, incorporating both DCT (Discrete Cosine Transform) and block DCT to compute global and block-level frequency information. Li et al. [11] proposed a novel approach involving a single center loss mechanism to learn features sensitive to frequency. Chen et al. [17] employed a two-stream architecture for forgery detection, with one stream processing RGB images and the other focused on high-frequency components. They segmented features into blocks to detect inconsistencies among these blocks. Chhabra et al. [18] exposed artifact information under different compression levels by learning a visibility matrix.

C. Transformer-Based Deepfake Detection

CNN architectures are highly effective at learning local features through mechanisms such as local receptive fields, shared weights, and spatial subsampling. However, the limited receptive fields of CNNs pose challenges in capturing global information. On the other hand, the Transformer architecture, initially designed for natural language processing tasks [10], has demonstrated its effectiveness in capturing long-range contextual information. In traditional Transformer models, this capability is achieved through self-attention mechanisms that model relationships between equally significant tokens. The adaptability of the Transformer has facilitated its application across various domains beyond text processing. For instance, in computer vision, the Vision Transformer (ViT) converts images into a series of flat tokens, which are then processed by a Transformer encoder for tasks like image classification. Similarly, the Multi-Scale Token Transformer (M2TR) [19] introduces a novel approach by splitting inputs into tokens of varying sizes, enabling the integration of multi-scale information within the Transformer framework for improved visual representation. Chen et al. [20] proposed a two-stage network tailored for face-related tasks. In the first stage, they employ a Transformer to learn both intra-class similarities and inter-class differences among faces. In the second stage, the network is fine-tuned based on the knowledge gained in the initial stage to perform classification tasks, leveraging the strengths of Transformer models for face-related applications.

D. Deepfake detection based on constrained learning

Some prior methods have explored the use of constraint learning to enhance model performance. Zhou et al. [21] introduced a two-stream network to detect inconsistencies in low-level facial features. Nirkin et al. [22] leveraged the inconsistency between the face and its contextual content for face forgery detection. Chen et al. [17] learned cosine inconsistency between feature blocks, subsequently using existing Masks to constrain the inconsistency. Li et al. [12] employed the two most dissimilar blocks between blocks for constraints, relying solely on image-level labels. Yang et al. [23] simulated the process of face forgery generation and tracked the underlying texture details during image generation.

In summary, existing works typically rely solely on either spatial domain features or frequency domain features, seldom considering their correlation and complementarity. This paper introduces an enhanced attention mechanism to realize frequency domain-guided spatial feature extraction instead of directly utilizing frequency domain features. This approach may help mitigate the potential impact of frequency domain features on model generalization. In contrast to the traditional Transformer, we posit that long-distance Token Mix may not be highly effective for forensic tasks. This is because artifact features are subtle, and most tokens represent content information, causing the interaction of numerous content information tokens to dilute the artifact information. This could adversely affect the detector's ability to extract artifacts. Recent studies [24] also support this perspective. Therefore, based on the Self-Attention (SA) concept, we devised a more refined frequency domain guidance module to instruct the spatial domain in learning subtle artifacts. Simultaneously, noise features are employed as a supplementary source, and the features learned in the spatial stream and noise stream are fused to enhance detection accuracy and generalization performance, especially when processing low-quality face images. To better distinguish among various forgery methods and improve accuracy, we enhanced the constrained loss proposed by Li et al. [12] and incorporated it into the second stage of our model training.

III. Feasibility Analysis

A. Fine-Grained Frequency Domain Clues

Forgery traces are typically displayed as high-frequency responses, and frequency domain information can effectively capture these clues. Previous studies have indicated that Fast Fourier Transform (FFT) or Discrete Cosine Transform (DCT) can be used to extract frequency domain-related forgery clues.

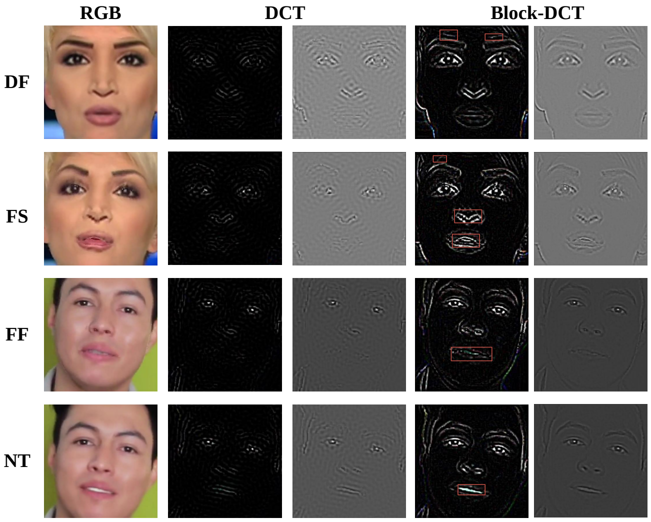


Fig. 1. Visual comparison of traditional DCT changes and Block-DCT changes.

However, most of these methods directly extract frequency domain features and merge them with those of the RGB. This results in coarse utilization of frequency domain information and difficulty in fully exploiting fine-grained forgery clues. To address these issues, we proposed a Double-Frequency Transformer Module (DFTM) approach to comprehensively mine fine-grained frequency domain features. We also utilized block discrete Fourier transform (Block-DCT) with sliding windows to capture fine-grained clues. This approach generates sufficient feature combinations, revealing potential artifacts that traditional frequency domain methods may overlook. Given common post-processing operations such as dataset compression, traditional frequency domain filtering methods can introduce noticeable ringing effects, which significantly impact artifact detection. However, our proposed Block-DCT methods effectively mitigates the ringing effect while mining the fine-grained clues.

At the same time, we conducted a visual analysis to elucidate the role of fine-grained clues in detecting fake faces and showcase the block-DCT method's effectiveness, as detailed in Figure 1. In this figure, the second and third columns display binary and grayscale images generated by traditional DCT transformations, while the fourth and fifth columns depict those generated by Block-DCT method. Considering the second and fourth columns, we observe that traditional DCT only filters out obvious high-frequency edge information, which is coarse and fails to reveal forgery traces in detail. In contrast, our Block-DCT approach employs an 8 × 8 block and filters on the image at a fixed step size, effectively uncovering detailed artifacts. Furthermore, examining the grayscale images, we noted that traditional DCT introduces noticeable ringing effects when filtering images subjected to post-processing operations (e.g., compression). These artifacts caused by the ringing effect can mislead the detection process. In contrast, our proposed Block-DCT exposes forgery traces while avoiding the ringing effect.

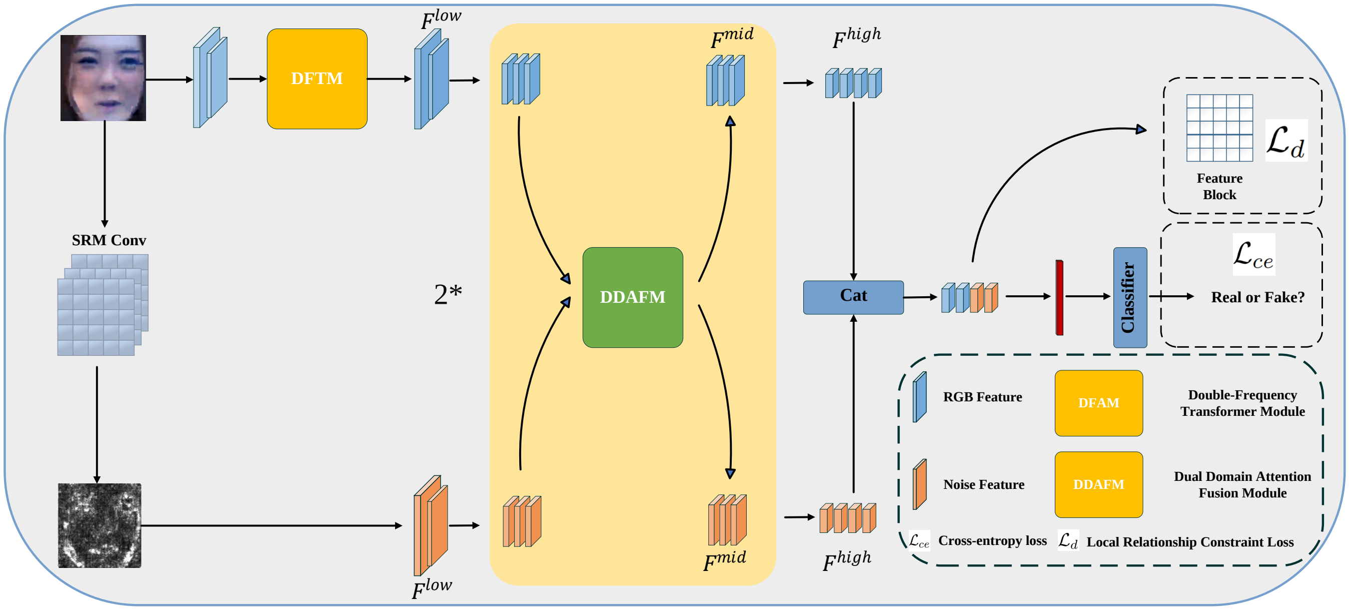


Fig. 2. Proposed model. It consists of a spatial flow guided by a Double-Frequency Transformer Module (DFTM) and a noise flow. The spatial flow is guided to focus on finer features, the noise flow mines the inconsistency of fake face noise, and finally combines the Dual Domain Attention Fusion Module (DDAFM) for feature interaction. Meanwhile, we use cross-entropy to train model in the first stage; the second stage uses cross-entropy loss to jointly train model with the proposed local correlation constraint loss.

B. Improving Self-Attention Mechanism

In high-quality datasets, traditional CNNs can reveal evident traces of forgery in the RGB domain. However, in highly compressed datasets, most forgery traces are faded by post-processing operations such as compression, leading to significant performance degradation in many existing methods. However, subtle clues that can be extracted remain despite the elimination of most forgery traces. Nevertheless, these subtle traces may disappear during forward network propagation, making their detection challenging. To address this issue, our approach aims to capture subtle forgery traces and enhance the information loss during forward propagation. Therefore, we propose leveraging the self-attention mechanism. While it is typically for downstream tasks in computer vision and effectively aggregates content features to enhance image representation, the traditional self-attention mechanism is not ideally suited for forensic tasks, which prioritize detailed forgery traces over a large number of content features. Thus, we optimized the self-attention algorithm to achieve better aggregate artifact information.

Specifically, the conventional self-attention mechanism is designed to aggregate regions with similar content information, which can be beneficial for various tasks. However, in the context of face forgery detection, where tampered areas are often small and the artifact information is subtle, directly applying the traditional self-attention mechanism may inadvertently focus on genuine regions rather than forged ones. This unintended aggregation of genuine content could obscure subtle forgery clues, ultimately impacting detection accuracy. Conversely, by facilitating the interaction among pixels within each token, our approach enhances regions with high-frequency artifact responses while leaving unaffected areas unchanged. This method effectively amplifies the representation of subtle forgery traces, particularly in highly compressed data, a capability that was previously lacking in the literature.

IV. PROPOSE APPROACH

A. Overview

Frame-level deepfake video detection is typically regarded as a binary classification problem. In this context, each frame of the video is considered an individual image with dimensions $C \times H \times W$, where C represents the number of image channels, and H and W denote the image's height and width, respectively. The objective is to design a model that provides a probability score, indicating the given frame's likelihood to be false or genuine. In this section, we present our framework in detail.

We introduce the Double-Frequency Transformer Module (DFTM), utilizing high-frequency information to guide spatial information, enabling a focus on more detailed artifacts. We employ SRM convolution to extract noise features, aiding in the detection of noise inconsistencies. Additionally, we propose the Dual Domain Attention Fusion Module (DDAFM) to combine these two complementary forgery clues. Finally, we detail the local relationship-constrained loss and training strategy. The entire network framework is depicted in Fig. 2.

B. Double-Frequency Transformer Module

As discussed earlier, the Self-Attention (SA) tends to aggregate features, enhancing image content features and long-distance dependence while diminishing high-frequency signals. However, forensic assignments, such as deepfakes detection, differ from other visual tasks because detailed artifacts are crucial for examining traces of forgery. Therefore, designing an SA mechanism that simultaneously enhances content and artifact information is paramount. Base on the concepts introduced in the F3Net and M2TR studies, this paper proposes a more refined frequency domain guidance module, as illustrated in Fig. 3.

Specifically, when processing an input image represented as $X \in \mathbb{R}^{(H \times W \times 3)}$, we initially employed a series of convolutional layers to extract relevant features from the image, thereby generating shallow features $F \in \mathbb{R}^{((\frac{H}{4}) \times (\frac{W}{4}) \times C)}$. These features were then fed into the DFTM module, where a Discrete Cosine Transform (DCT) and a high-pass filter were applied to the F feature. Unlike conventional high-pass filters, we employed block high-pass filtering on the features. This approach was selected due to normal high-pass filters generating ringing effects in compressed or recompressed images, adversely affecting detector performance. Drawing inspiration from F3Net, we manually designd a filter (mask) f_{base}^{i} in the DFTM module, which filter partitions the frequency band into high and low-frequency bands. As a result, low-frequency components were filtered out. Additionally, we introduced a learnable filter f_{w}^{i} to the mask. The specific formula is as follows:

$$F_H = D^{-1}[D(F) \odot [f_{base}^{i} + \sigma(f_{w}^{i})]] \tag{1}$$

In the provided equations, D and D^{-1} represent the Discrete Cosine Transform (DCT) and its inverse, $\odot$ denotes the dot product, and $\sigma(x) = \frac{1-\exp(-x)}{1+\exp(-x)}$ compresses x within the range of -1 to 1. In this context, i denotes the filtered band, with the experiment focusing solely on the high-frequency band. After filtering the F feature with the specified filter, the high-frequency feature F_H was derived. Subsequently, F_H was utilized to guide the spatial feature F. This process is divided into two modules: the Frequency Spatial Fusion Module (FSFM) and the Frequency Enhance Module (FEM), which incorporate the designed Local Enhance Module (LEM) framework.

1) Local Enhance Module (LEM): To facilitate more intricate artifact features detection, we utilized the frequency domain to guide the learning process within the spatial domain. We hypothesize that artifact information is closely associated with the surrounding local details. Instead of employing a traditional token mix, we introduced a Local Information Enhancement Module (LEM). Initially, we segmented the feature F into blocks according to patch size, generating several tokens. Notably, no interaction between these tokens exists; instead, self-interaction occurs among the pixels within each token. Therefore, the primary objective of the local information enhancement module is to amplify local artifact and content information, ensuring an evident demarcation between genuine and false feature regions. The specific formula is expressed as follows:

$$Q, K, V = Conv(F) \tag{2}$$

$$Q, K, V \in R^{h \times w \times c} \rightarrow Q_1, K_1, V_1 \in R^{(\frac{h}{G} \times \frac{w}{G}) \times (G \times G) \times c} \tag{3}$$

$$O = Softmax\left(\frac{Q_1 \otimes {K_1}^T}{\sqrt{d_k}}\right) \otimes V_1 \tag{4}$$

$$O \in R^{(\frac{h}{G} \times \frac{w}{G}) \times (G \times G) \times c} \rightarrow O \in R^{h \times w \times c} \tag{5}$$

Where $Conv$ represents a vanilla 1×1 convolution, $\otimes$ denotes matrix multiplication, G represents the set Patch Size, and $\sqrt{d_k}$ scales the Attention to prevent the vector dot product from becoming excessively substantial. The local enhancement module is depicted in Fig. 4.

Furthermore, we enhanced the artifact information at various scales. Specifically, we partitioned the input feature map F into four features $F_i \in R^{(\frac{H}{4}) \times (\frac{W}{4}) \times (\frac{C}{4})}$ $i = 1, 2, 3, 4$. Under different feature blocks F_i, we employed distinct Patch sizes for local enhancement, thereby enhancing local artifact information across the various scales. After that, we set the Patch Size to 28 × 28, 14 × 14, 7 × 7, and 4 × 4. This calculation approach reduces computational complexity. Then, we selected the $Patch_{size}$ and $Head_{num}$ values according to the traditional Transformer approach. By setting the number of $Head_{num}$ to 4, the $Patch_{size}$ under different heads becomes $H/8$, $H/16$, $H/32$, and $H/64$ respectively. This configuration enables the DFTM module to globally model and locally enhance fine-grained clues, further reducing the computational complexity.

2) Frequency Enhance Module (FEM): We designed a frequency domain enhancement module to facilitate subtle artifact learning within the spatial domain, utilizing high-frequency information to guide spatial learning. Specifically, we used high-frequency information as Q and K, while spatial features served as V. These components were then input into the Local Enhance Module (LEM), as described below:

$$F_{FE} = LEM(F, F_H) \tag{6}$$

$$Q, K = Conv\left(F_H\right) \;\; V = Conv(F) \tag{7}$$

Get the frequency domain guidance feature F_{FE}.

3) Frequency Spatial Fusion Module (FSFM): Relying exclusively on frequency domain guidance may result in the loss of crucial spatial information, which is vital for detection. To address this challenge, we introduced a fusion module that utilizes frequency domain information to complement the missing spatial data without generating redundant content. The implementation details are similar to the Frequency Enhance Module (FEM); however, the main difference is the utilization of high-frequency information as Q, while the spatial features serve as K and V. The resulting Q, K, and V were then input into the Local Enhance Module (LEM), as shown below:

$$F_{FS} = LEM(F, F_H) \tag{8}$$

$$Q = Conv\left(F_H\right) \;\; K, V = Conv(F) \tag{9}$$

Get the fusion feature F_{FS}.

Finally, the features F_{FE} and F_{FS} were concatenated along the channel dimension using the Concat operation. Subsequently, we performed dimensionality reduction and feature aggregation using convolution to obtain the feature F_{DF}. The specific expression is as follows:

$$F_c = Concat(F_{FE}, F_{FS}) \tag{10}$$

This article has been accepted for publication in IEEE Transactions on Artificial Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAI.2024.3455311

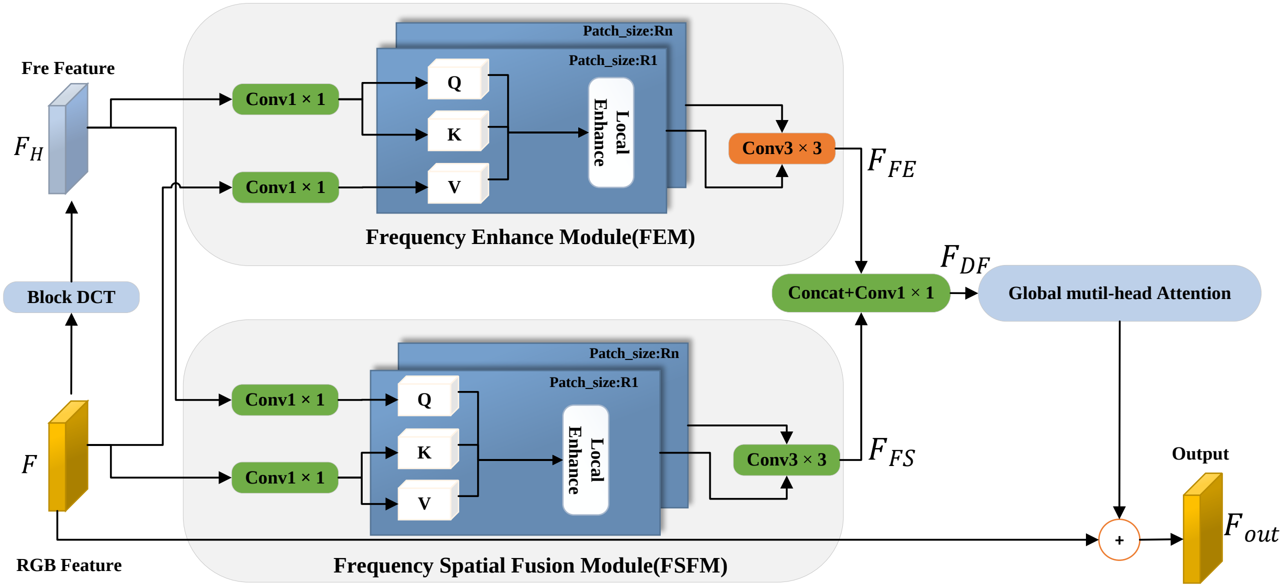


Fig. 3. Double-Frequency Transformer Module (DFTM). The spatial features are transformed by block DCT and high-frequency filtering to obtain the frequency domain features, and the frequency domain features are used to guide and fuse the spatial features, and finally a coarse-grained SA is used to obtain more refined features.

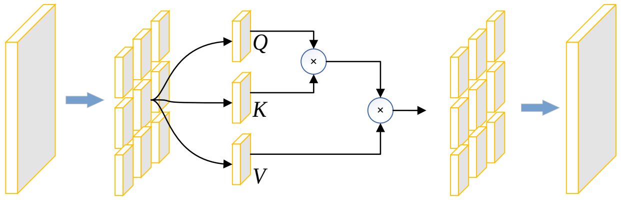
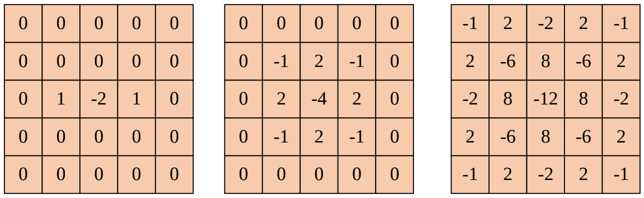


Fig. 4. Local Enhance Module.



0	0	0	0	0
0	0	0	0	0
0	1	-2	1	0
0	0	0	0	0
0	0	0	0	0

0	0	0	0	0
0	-1	2	-1	0
0	2	-4	2	0
0	-1	2	-1	0
0	0	0	0	0

-1	2	-2	2	-1
2	-6	8	-6	2
-2	8	-12	8	-2
2	-6	8	-6	2
-1	2	-2	2	-1

Fig. 5. SRM convolution kernel.

$$F_{DF} = BN(Conv_2(\sigma(BN(Conv_1(F_c))))) \tag{11}$$

Where $Conv_1$ represents vanilla 1×1 convolution, which reduces the feature dimensionality. BN stands for Batch Norm, σ represents the ReLU activation function, and $Conv_2$ denotes the vanilla 3×3 convolution, which aggregates features.

At this stage, we assume the feature has effectively distinguished between the genuine and false regions. Therefore, coarse-grained Self-Attention (SA) can be utilized to model the features over long distances. Specifically, we downsampled the feature F_{DF} by a factor of seven using standard convolution. Then, traditional SA was performed on the downsampled features for global modeling. Finally, a residual connection was established with the input feature F, resulting in the output feature F_{out}. This process is expressed as follows:

$$F_{out} = SA\left(Conv_{down}\left(F_{DF}\right)\right) + F \tag{12}$$

C. Noise Stream

We use the original SRM convolution to extract noise flow features for face forgery detection through noise inconsistency. SRM uses a high-pass filter to model the noise residual $R_{i,j}$:

$$R_{i,j} = P\left(\mathcal{N}_{i,j}\right) - I_{i,j} \tag{13}$$

Here, $\mathcal{N}i,j$ represents the local neighborhood surrounding the central pixel Ii,j. The function $P()$ computes the prediction for $I_{i,j}$ based on the information within this neighborhood, typically implemented using a standard convolution operation. The residuals are then quantified with q, rounded and truncated at the end:

$$R_{i,j} \leftarrow trun(rount(\frac{R_{i,j}}{q})) \tag{14}$$

We selected 3 commonly used filters among 30 SRM filters. As shown in Fig. 5.

D. Dual Domain Attention Fusion Module

Attention mechanisms have benn extensively applied in natural language processing and computer vision domains. For cross-modal attention, GFFD [25] and F3Net are used to design a dual-modal attention fusion module based on spatial correlation. Thus, we employed F3Net's Mixblock as our spatial fusion module to model the interactions between spatial information and high-frequency noise within the spatial locations. We also introduced an innovative channel attention mechanism designed to weigh the significance of local features

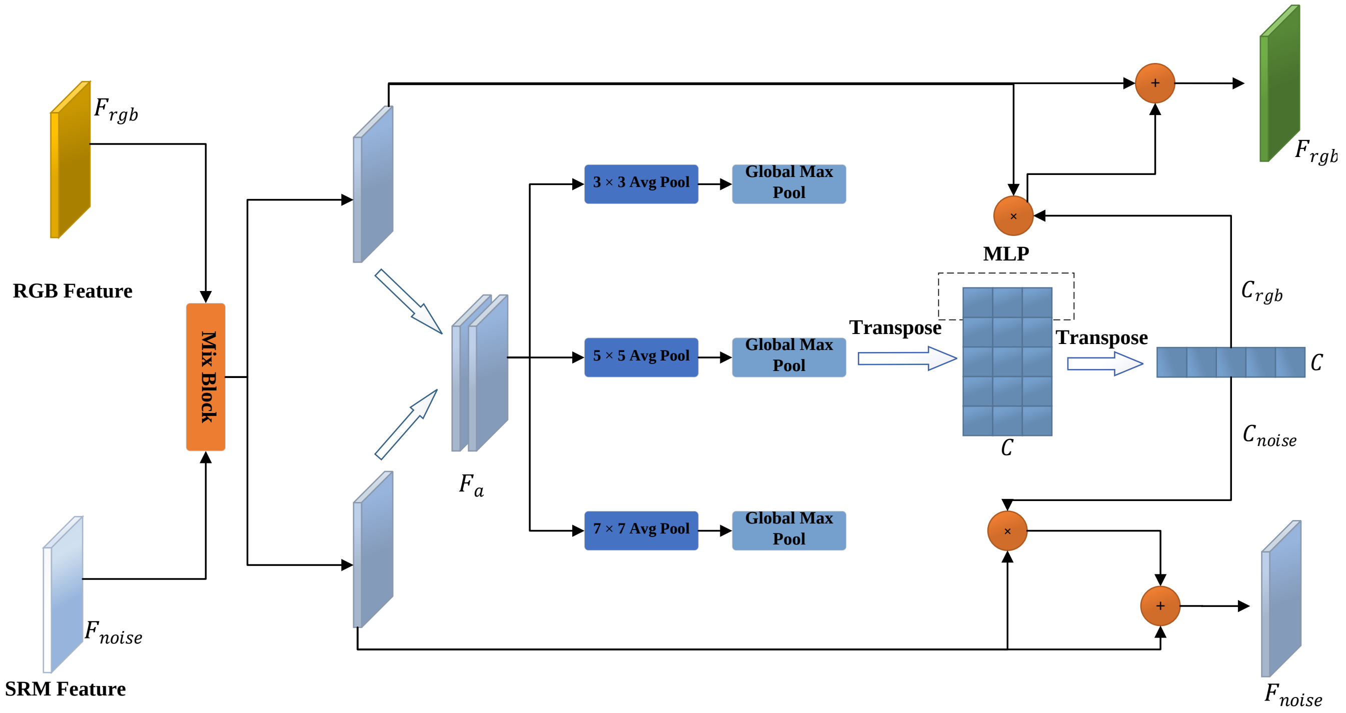


Fig. 6. Dual Domain Attention Fusion Module (DDAFM). First, the spatial features are fused through MixBlock modeling, and then the channel features are enhanced through the multi-scale channel attention we designed.

while simultaneously mitigating the impact of background noise. Fig. 6 shows our designed dual domain attention fusion module. As shown in the figure, we used MixBlock to model the interaction between the spatial and noise features in the location space. Then, we use ourd proposed channel attention mechanism to mine artifacts and reduce the noise effects at different scales. The formula is described as follows:

$$F_a = Concat(F_{rgb}, F_{noise}) \tag{15}$$

We used 3×3, 5×5, and 7×7 average pooling for F_a, and global max pooling to obtain the three features C_i:

$$C_i = GMP\left(AvgPool_{ixi}\left(F_a\right)\right), i = 3,5,7. C_i \in R^{1\times1\times2C} \tag{16}$$

Next, we concatenated the features along the first dimension, transpose the resulting tensor, and utilized an MLP to aggregate features of different scales. Finally, we applied the Sigmoid activation function to achieve channel attention. The specific formula is expressed as:

$$C = transpose\left(Concat\left(C_i\right)\right), C \in R^{1\times2C\times3} \tag{17}$$

$$C = \sigma(traspose(MLP\left(C\right))), C \in R^{1\times1\times2C} \tag{18}$$

Additionally, the tensor was split along the channel dimension to obtain channel attention for both the spatial and noise flows. Then, these channel attentions were connected with the original features using the dot product and residual connection to obtain the output features, as described below:

$$C_{rgb}, C_{noise} = chunk\left(C\right) \tag{19}$$

$$F_{rgb} = F_{rgb} * C_{rgb} + F_{rgb} \tag{20}$$

$$F_{noise} = F_{noise} * C_{noise} + F_{noise} \tag{21}$$

E. Local Relationship Constraint Loss

We integrated information from various scales to assess the inconsistencies within the local regions based on learned spatial and noise information. Specifically, we combined information from the shallow, medium, and deep network layers using the following formula:

$$F^l = Conv(F^l_{rgb}) \ + \ Conv(F^l_{noise}) \tag{22}$$

In this formula, F^l denotes the fused feature of layer l, and $Conv$ represents a standard 1×1 convolution. Additionally, we incorporated multi-scale information, leveraging high-resolution low-level features for localization and low-resolution high-level semantic information for recognition. Precisely, we resized the features F^{low}, F^{mid}, and F^{high} from various scales into a uniform scale; then, we concatenated them along the channel dimension to obtain the multi-scale feature F^{all}.

To delineate the inconsistencies within the local regions, we partitioned the feature F^{all} into blocks of size $p \times p$ and a stride of s, with each block denoted as $\widetilde{u}$. The resulting $p \times p$-sized blocks $\widetilde{u}$ were flattened into one-dimensional vectors. Thus, the inconsistencies between each block and others were computed based on the cosine distance, as expressed below:

$$s_{x,z} = 1 - \left\langle \frac{\widetilde{u}_x}{\|\widetilde{u}_x\|} | \frac{\widetilde{u}_z}{\|\widetilde{u}_z\|} \right\rangle \tag{23}$$

$s_{x,z}$ represents the uncorrelation between block $\widetilde{u}_x$ and $\widetilde{u}_z$.

Moreover, we identified certain shortcomings in Li et al.'s proposed [12] approach. Focusing solely on the two blocks with the greatest discrepancies may render the network susceptible to background noise, leading it to prioritize irrelevant noise features. Hence, we introduced enhancements to address this issue. Assuming there are n blocks, we computed the average distance s_i^{mean} between each block $\widetilde{u}_i$ and the others by averaging their uncorrelation. Subsequently, we obtaine the average inconsistency matrix $s_n^{mean} \in R^{n\times 1}$ between the blocks. To ensure that the constraint loss is not influenced by edge noise, we calculated the weighted average for the average distances of the first N blocks to derive the local maximum difference value M. This process is expressed as follows:

$$s_i^{mean} = Mean\left(\sum s_{ij}\right)\ j \in 1, 2, \ldots, n \tag{24}$$

$$M = Mean\left(\sum_0^N Topk\left(s_n^{mean}, k\right)\right) \tag{25}$$

Where $Mean$ denotes the average value computation, and $Topk$ represents the first k maximum values selection. Notably, owing to the presence of face and authentic backgrounds, the discrepancies in local features among false face images are significantly more pronounced than those observed in genuine face images:

$$M_f > M_r \tag{26}$$

In this instance, M_f and M_r correspond to the M values for the false and the real face images, respectively. Consequently, we employed the following constrained loss function to facilitate the inconsistency learning:

$$\mathcal{L}_d = max(0, 1 - \frac{1}{|\Omega_f|}\sum M_f + \frac{1}{|\Omega_r|}\sum M_r) \tag{27}$$

Where Ω_f and Ω_r represent the collection of false and genuine faces, respectively. We also used $\mathcal{L}_d$ for training in the network's second stage. During the experiment, we observed that jointly optimizing the network with two losses did not yield satisfactory detection performance. We speculate that this is due to the local relationship constraint loss's focus on the variations within the cluster class, while the cross-entropy loss emphasizes the differences between classes. As a result, the model may struggle to find a balance between the two losses during the training process. This issue can be effectively addressed through a two-stage learning approach. In the first stage, we solely employed the cross-entropy loss to model the differences between classes and establish the fuzzy margins. In the second stage, we utilized the local relationship constraint and cross-entropy losses for joint optimization. This stabilizes the fuzzy boundaries between classes using cross-entropy loss. At the same time, it aggregates the feature distribution of the fuzzy boundary at the closest class center point through the local relationship constraint loss, thereby improving the results. Hence, our proposed method has two loss stages of losses respectively:

The initial stage employs cross-entropy loss, enabling the network to categorize the data:

$$\mathcal{L}_{total} = \mathcal{L}_{ce} \tag{28}$$

In the second stage, we train the network using a combination of the cross-entropy and local relationship constraint losses, as expressed below:

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \lambda\mathcal{L}_d \tag{29}$$

In this case, λ is a hyperparameter that regulates the trade-off between the cross-entropy and supervised contrastive losses. In addition, λ was set to 0.5 in our empirical settings.

V. Experiments

In this section, we explain the model's implementation specifications and provide visualizations of the results. Then, we conduct experiments to evaluate our proposed method's effectiveness. Comparative assessments against state-of-the-art techniques are also performed to assess its performance.

A. Data Preparation

FaceForensics++ (FF++) [3]: This dataset is widely employed in the face forgery detection field. It comprises 1,000 authentic and 4,000 manipulated videos. The manipulated videos include four common tampering techniques: Deepfakes, Face2Face, FaceSwap, and NeuralTextures. Furthermore, the dataset offers three video compression variants: raw, C23, and C40. We adhered to official specifications while constructing our dataset. We allocated 720 videos to training and reserved 140 for validation and testing. Then, we sampled 80 frames from each video and extracted 20 frames for verification and testing. To ensure a balanced real-to-fake data ratio, we employed a quadruple upsampling strategy for the genuine data. In addition, our experimental evaluations were conducted on both the C23 and C40 versions of the dataset.

Celeb-DF [38]: Celeb-DF is a recently introduced and challenging dataset centered on deepfake videos. It encompasses 590 authentic and 5639 fake videos precisely generated using sophisticated deepfake algorithms from YouTube. We utilized the official Celeb-DF test set for our evalutaion.Then, we extracted 32 frames from each video within this set to comprehensively assess the model's generalization performance across diverse databases.

DFDC [39]: is a notable dataset in the face swap video realm, comprising a vast collection of over 100,000 video clips. These clips feature the participation of 3,426 paid actors and encompass a wide spectrum of production techniques, including Deepfake, GAN-based, and non-learned methods.

WildDeepfake [32]: is a comprehensive dataset featuring 7,314 facial imagery sequences. These sequences are exclusively collected from the internet and are partitioned into 6,508 samples for training and 806 for testing purposes. A notable characteristic of this dataset is its inherent diversity, spanning various aspects, including activities, settings, and manipulation techniques. This diversity the dataset's complexity level and aligns it more closely with real-world face manipulation scenarios.

FachShifter [40]: is an efficient face-changing model created using the dual-stage GAN method and has enhanced visual effects and fidelity.

TABLE I
HERE ARE THE COMPARISON RESULTS ACROSS THE FF++ (C23), FF++ (C40), FSH (C23), FSH (C40), AND WILDDEEPFAKE DATASETS.

Methods	**FF++ (C23)**		**FF++ (C40)**		**Fsh (C23)**		**Fsh (C40)**		**WildDeepfake**	
	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC
MesoNet [1]	60.51%	74.55%	61.24%	71.41%	70.20%	88.52%	57.06%	73.42%	35.61%	69.53%
MesoNet-Inc4 [1]	82.15%	83.64%	54.38%	74.28%	88.40%	99.07%	82.41%	95.44%	65.74%	81.79%
Xception [26]	89.84%	98.14%	84.64%	87.60%	96.64%	99.75%	93.56%	98.17%	82.55%	89.94%
EfficientNetb4 [27]	91.89%	98.45%	84.55%	88.71%	97.41%	99.76%	93.36%	98.53%	80.31%	89.96%
F3-Net [6]	93.78%	98.55%	83.89%	88.48%	97.24%	99.74%	92.16%	98.40%	80.68%	88.29%
RFM [28]	91.59%	98.37%	84.42%	87.55%	98.03%	99.70%	93.16%	98.46%	80.34%	88.96%
GRAM [29]	92.21%	97.81%	83.31%	85.44%	97.51%	99.62%	92.80%	97.59%	80.11%	88.22%
GFFD [25]	90.23%	98.28%	84.25%	87.71%	96.09%	99.70%	91.81%	98.23%	80.56%	90.10%
SPSL [30]	91.50%	95.32%	81.57%	82.82%	-	-	-	-	-	-
M2TR [19]	94.08%	98.43%	85.11%	89.23%	97.80%	99.65%	92.69%	98.38%	79.96%	87.97%
HFI-Net [31]	91.87%	97.07%	85.69%	88.40%	-	-	-	-	-	-
Add-Net [32]	93.04%	98.34%	83.20%	87.89%	-	-	-	-	76.25%	86.17%
MaDD [33]	95.37%	98.66%	84.95%	87.26%	-	-	-	-	82.86%	90.71%
GocNet [34]	91.67%	97.58%	83.15%	85.50%	97.10%	99.54%	90.07%	96.73%	81.19%	88.36%
F2-Trans [35]	**96.60%**	**99.24%**	**87.20%**	89.91%	-	-	-	-	-	-
SFIC [36]	91.86%	97.91%	83.69%	85.73%	91.95%	98.70%	88.26%	96.70%	75.91%	83.55%
MFFLE [37]	95.49%	98.84%	85.75%	89.49%	95.45%	99.66%	93.14%	98.57%	78.29%	90.60%
Ours	95.20%	98.68%	85.78%	**89.98%**	**98.07%**	**99.77%**	**93.74%**	**98.78%**	**83.71%**	**91.22%**

B. Implementation Details

In the preprocessing step, we utilized the open-source DLIB detector [41] to detect and align faces within all the frames extracted from the unaltered and manipulated videos. Each image was subsequently resized to a dimension size of 256×256 and subjected to random cropping, yielding a final input size of 224×224 for the network. Our proposed method was implemented using PyTorch [42] and trained on an NVIDIA Tesla A30 GPU. For feature learning, we utilized EfficientNet-b4 [27], which was pre-trained on ImageNet. We removed the fully-connected layers within this configuration and opted to use the widespread cross-entropy loss as our first-stage loss function. Then, we optimized this loss using the AdamW optimizer [43] with the initial learning rate set at $2e-4$ and a weight decay of $1e-2$. The Batch Size was set to 12 and the model was trained for 20 epochs. In the second stage, the cross-entropy and local relationship constraint losses were used to jointly optimize the model with the best performance occurring in the first stage, where N is 8. The same optimizer and parameters as the first stage were used to continue training the model for 20 epochs.

C. Performance Evaluation

To demonstrate our approach's effectiveness, we conducted a comprehensive comparison with various representative methods. Our selection included previous well-established methods and a range of recent, outstanding, and open-source approaches.

1) Effectiveness of Intra-Dataset Detection: Our comparative analysis encompasses state-of-the-art methods on multiple datasets, including FF++ (C23), FF++ (C40), Fsh (C23), Fsh (C40), and WildDeepfake. We also utilized Accuracy (Acc) and Area Under Curve (AUC) as evaluation metrics. The comparative results are presented in Tab. I. In addition, we compared our approach with many recent methods, such as MesoNet [1], Xception [26], RFM [28], F3Net [6], GFFD [25], GRAM [29], Add-Net [32], MaDD [33], M2TR [19], HFI-Net [31], GocNet [34], F2-Trans [35], SCFI [36], and MFFLE [37]. The experimental results show that F2-Trans performs excellently on both the FF++ (C23) and FF++ (C40) datasets. Additionally, MFFLE outperforms our method on the FF++ (C23) dataset. While our method surpasses recently published works on some datasets, it performs slightly less effectively on the FF++ (C23) dataset. We speculate that our method may overly emphasize detailed forgery clues while potentially overlooking the contribution of content features to detection. Moreover, both the MaDD [33] and MFFLE [37] methods performed well on the WildDeepfake dataset; however, our method still achieves the best performance in comparison. Our method performed well on multiple datasets, particularly in FF++(C40) where the data underwent significant compression. Many methods struggle to capture relevant artifacts in highly compressed scenarios. However, our approach exhibits robustness under such conditions. Our strategy's resilience arises from utilizing high-frequency information to guide spatial awareness within the model, enabling it to focus on fine-grained artifacts while filtering out redundant content. Additionally, we employed noise inconsistency features as an auxiliary signal, offering a multi-faceted approach to detection. Therefore, we effectively demonstrated our method's proficiency in addressing challenges posed by heavily compressed contexts through experimentation.

TABLE II
WE PRESENT QUANTITATIVE COMPARISON RESULTS FOR CROSS-MANIPULATION EXPERIMENTS CONDUCTED ON THE FF++ DATASET UNDER THE C23 SETTING.

Training Set	**Methods**	**Testing Set (AUC)**				**Avg**
		DF	**F2F**	**FS**	**NT**	
DF	GRAM [29]	99.58%	66.67%	27.65%	73.55%	66.86%
	GFFD [25]	99.56%	64.52%	31.44%	74.56%	67.52%
	M2TR [19]	99.57%	**67.07%**	26.82%	75.33%	67.20%
	Ours	**99.59%**	65.85%	**31.66%**	**77.35%**	**68.61%**
F2F	GRAM [29]	**83.58%**	99.13%	40.62%	**72.64%**	74.00%
	GFFD [25]	78.23%	99.20%	58.70%	65.28%	75.35%
	M2TR [19]	74.33%	99.25%	54.64%	58.67%	71.72%
	Ours	83.26%	**99.46%**	**65.13%**	69.38%	**79.31%**
FS	GRAM [29]	56.30%	60.79%	99.50%	**51.66%**	67.06%
	GFFD [25]	55.65%	66.16%	**99.61%**	50.23%	67.93%
	M2TR [19]	56.16%	66.13%	99.60%	50.56%	68.11%
	Ours	**56.55%**	**70.24%**	99.53%	50.76%	**69.27%**
NT	GRAM [29]	92.40%	**73.57%**	39.93%	97.08%	**75.75%**
	GFFD [25]	91.96%	70.05%	34.37%	97.55%	73.48%
	M2TR [19]	90.42%	68.84%	**44.78%**	97.49%	75.39%
	Ours	**92.59%**	69.38%	34.00%	**97.60%**	74.11%

2) Extending the Scope of Cross-Manipulation Detection: Assessing a method's generalization ability plays a pivotal role in deepfake detection. To gauge our method's generalization capabilities, we conducted cross-detection experiments on the FF++ (C23) dataset. In this instance, our model was trained using fake images from one specific method and tested with those from four different methods. We compared our results with those of three state-of-the-art methods, summarizing the outcomes in Tab. II. The results emphasize our method's robust generalization capacity across distinct subsets of the same dataset, demonstrating its effectiveness in detecting deepfakes from various sources. However, our method's generalization effect on a single subset of neural textures is less robust than other methods. This may be due to the fine nature of the artifacts in this particular subset. Furthermore, when the model is trained using only this subset, it tends to overfit its nuances, leading to suboptimal generalization on other subsets. Notably, our approach and the GFFD method performed relatively better on this single subset, albeit with insufficient generalization. Conversely, the GRAM method demonstrates poor performance on this subset, suggesting a lack of learned unique operation traces specific to this subset, resulting in improved generalization.

TABLE III
COMPARATIVE RESULTS ON CELEB-DF, DFDC, AND WILDDEEPFAKE DATASETS USING FF++ (C23) AS THE TRAINING DATASET.

Methods	**Train Set**	**Test Set (AUC)**		
		Celeb-DF	**DFDC**	**WildDeepfake**
MesoNet-Inception4 [1]	**FF++ (C23)**	56.82%	54.32%	62.85%
Xception [26]		63.27%	66.84%	70.41%
EfficientNetb4 [27]		65.63%	66.60%	70.73%
RFM (Xception) [28]		69.19%	67.33%	72.99%
GRAM [29]		70.63%	65.97%	70.05%
GFFD [25]		64.36%	67.90%	73.49%
M2TR [19]		66.04%	69.00%	71.69%
MFFLE [37]		69.84%	**71.28%**	73.04%
Ours		**72.76%**	69.57%	**74.51%**

3) Generalization Evaluation On Cross-dataset: To comprehensively assess our model's generalization performance in real-world scenarios, we extended our evaluations to cross-dataset experiments. Specifically, we trained our models using data from four forgery methods within the FF++ (C23) dataset and evaluated them on high-quality datasets such as Celeb-DF, DFDC, and WildDeepfake. The results presented in Tab. III demonstrate our method's superiority in generalization performance. This proficiency extends across a spectrum of manipulation techniques, including facial expression alterations and face replacements. When spatial information is insufficient for identifying artifacts, our approach effectively explores forgery clues using the noise inconsistency strategy.

Moreover, our method is mainly dedicated to detection in real life scenarios, and solves the robustness issues caused by the various post-processing operations of media images. While our method demonstrates good generalization performance compared to some current mainstream approaches, it still falls short when compared with existing methods [44]–[46] that specifically prioritize generalization ability. In the next stage, we will focus on improving generalization performance, building upon the foundations laid by existing methods.

TABLE IV
TRAIN ON THE ORIGINAL TRAINING SETS OF FF++ (C23) AND WILDDEEPFAKE, AND PERTURB THE TEST SET BY APPLYING JPEG COMPRESSION WITH DIFFERENT QUALITY FACTORS (QF) DURING EVALUATION.

Datasets	**Methods**	**QF = 100**	**QF = 85**	**QF = 70**	**QF = 55**	**QF = 40**
		AUC	**AUC**	**AUC**	**AUC**	**AUC**
FF++(C23)	F3-Net [6]	98.55%	97.49%	93.64%	85.43%	75.01
	GFFD [25]	98.30%	97.00%	92.73%	86.04%	77.46%
	M2TR [19]	98.43%	96.79%	92.17%	85.23%	78.54%
	Ours	**98.68%**	**97.68%**	**93.94%**	**86.72%**	**79.10%**
WildDeepfake	F3-Net [6]	88.29%	87.29%	86.76%	85.84%	85.03%
	GFFD [25]	90.10%	89.59%	88.89%	87.50%	85.40%
	M2TR [19]	87.97%	87.34%	86.95%	86.26%	84.81%
	Ours	**91.22%**	**90.63%**	**89.35%**	**87.82%**	**85.62%**

4) The Robustness Experiment: During the video capturing and transmission process, various noise types such as blur and salt were incorporated into the data. Therefore, the model's detection sensitivity to disturbances is essential for real-world applications. Hence, we validated our method's robustness through experimentation.

We introduced compression parameters that were not present during the training phase for testing purposes. Specif-

TABLE V
ROBUSTNESS EVALUATION UNDER VARIOUS TYPES OF PERTURBATIONS.

Datasets	Methods	+GaussianNoise		+SaltPepperNoise		+GaussianBlur	
		△Acc↓	△AUC↓	△Acc↓	△AUC↓	△Acc↓	△AUC↓
FF (C40)	F3Net [6]	4.84%	0.72%	5.05%	7.75%	1.95%	1.22%
	GFFD [25]	2.13%	1.87%	4.89%	6.91%	1.60%	1.20%
	M2TR [19]	2.47%	2.08%	3.70%	6.10%	**0.51%**	1.41%
	Ours	**0.50%**	**0.08%**	**2.72%**	**3.22%**	0.59%	**1.18%**
WildDF	F3Net [6]	4.77%	1.27%	4.49%	4.42%	3.91%	**0.28%**
	GFFD [25]	1.71%	1.28%	5.40%	7.15%	2.73%	0.59%
	M2TR [19]	1.88%	1.44%	5.50%	4.41%	3.27%	0.29%
	Ours	**1.57%**	**0.98%**	**3.04%**	**2.74%**	**1.28%**	0.46%

TABLE VI
ABLATION RESEARCH OF SEVERAL DESIGNED MODULES ON FF++ (C40).

Models	AUC
Efficient Net [27]	88.71%
Eff + DFTM	89.47%
Eff + DDAFM	89.04%
Eff + DFTM + DDAFM	89.76%
Eff + DFTM + DDAFM + Constraint Loss	**89.98%**

ically, we trained our model on the FF++ (C23) and WildDeepfake datasets; then, we subjected the test data to JPEG compression using different quality factors. The experimental results are presented in Table IV. Moreover, recognizing the importance of adapting forgery detection methods to real-world scenarios, we also conducted robustness experiments involving various post-processing operations. Accordingly, we applied Gaussian noise, salt and pepper noise, and Gaussian blur to the test data and measured the decay in Acc and AUC (denoted as △Acc and △AUC) to evaluate the model's detection robustness, as shown in the Table V. The results indicate that our method's performance slightly decreases in most cases.

D. Ablation Study

1) Ablation experiments of various modules of the model: To evaluate the efficacy of our framework's components (i.e., DFTM, DDAFM, and the relationship constraint loss), we conducted comprehensive ablation studies on the FF++ (C40) dataset scenes. We aim to examine each component's contribution to the improvement of overall performance. For a quantitative assessment, we evaluated our framework and its variants, and the results are presented in Table VI. On the FF++ (C40) dataset, we observed a significant 1.27% backbone improvement in AUC, underscoring the proposed module's efficacy. In addition, we anticipate that the two-stage training local relationship constraint loss will yield further performance enhancements. Furthermore, our approach demonstrates a substantial performance improvement. in comparison to other methods, such as GFFD and M2TR, with input sizes of 256 and 320, respectively. We observed marginal bachbone AUC improvements of 0.11% and 0.52%. In contrast, our method achieved a significant 1.27% improvement with a smaller image size of 224, exhibiting superior performance with fewer image features.

TABLE VII
ABLATION RESEARCH OF DFTM MODULES ON FF++ (C40).

Models	AUC
Efficient Net [27]	88.71%
Efficient Net + Self-Attention	85.92%
Efficient Net + FSFM	88.67%
Efficient Net + FEM	88.82%
Efficient Net + FSFM + FEM	**89.76%**

2) DFTM and DDAFM ablation experiments: Simultaneously, we conducted comprehensive ablation studies on small modules within different components. Initially, we performed ablation experiments on two small modules in DFTM to validate our hypothesis. The traditional Transformer self-attention, as employed in the DFTM module, predominantly enhances content information, which may not be conducive to forgery detection. The experimental results are detailed in Table VII.

Furthermore, we extended our ablation experiments to specifically focus on the multi-scale module within DFTM. This approach aimed to verify the multi-scale and Local Enhance method's effectiveness. The ablation experiments results are detailed in Table VIII.

TABLE VIII
ABLATION STUDIES AT DIFFERENT SCALES ON FF++ (C40).

AUC Methods Patch Size	Token Mix	Local Enhance
28 × 28	88.23%	89.30%
14 × 14	87.97%	89.14%
7 × 7	88.19%	88.90%
Mutil-Size	87.85%	89.76%

We conducted additional ablation experiments to assess the proposed multi-scale channel attention mechanism's effectiveness. The experimental results are presented in Table IX.

TABLE IX
ABLATION STUDIES AT DIFFERENT FUSION ON FF++ (C40).

Methods	AUC
Mix Block	88.96%
Mix Block + MSCA	**89.76%**

3) Local Relationship Constraint Loss ablation experiments: We conducted an ablation study comparing the original local constraint loss with our improved one. We employed two blocking methods for the improved loss:

- The block and stride sizes were set to 3;
- The block and stride size were set to 5 and 2, respectively.

Then, we conducted comprehensive comparative experiments under various conditions by setting N to 1, 4, 8, and

TABLE X
PERFORMANCE OF IMPROVED LOCAL CONSTRAINT LOSS IN DATASETS FF++ (C23) AND FF++ (C40). K REPRESENTS THE SIZE OF THE BLOCK, S REPRESENTS THE STEP SIZE, AND N REPRESENTS THE NUMBER OF BLOCKS.

Methods	FF++ (C23)		FF++ (C40)	
	ACC	AUC	ACC	AUC
Model	93.65%	98.38%	85.03%	89.76%
Model + LRCL	94.19% (+0.54)	98.67% (+0.29)	85.25% (+0.22)	89.87% (+0.11)
Model + LRCL (Ours, K = 3, S = 3, N = 1)	95.06% (+1.41)	98.65% (+0.27)	85.41% (+0.38)	90.00% (+0.24)
Model + LRCL (Ours, K = 3, S = 3, N = 4)	94.08% (+0.43)	98.65% (+0.27)	84.93% (-0.1)	89.80% (+0.04)
Model + LRCL (Ours, K = 3, S = 3, N = 8)	95.08% (+1.43)	98.60% (+0.22)	85.12% (+0.09)	89.83% (+0.09)
Model + LRCL (Ours, K = 3, S = 3, N = 16)	93.89% (+0.24)	98.64% (+0.26)	85.07% (+0.04)	89.83% (+0.09)
Model + LRCL (Ours, K = 5, S = 2, N = 1)	94.16% (+0.51)	98.70% (+0.32)	85.57% (**+0.54**)	89.98% (+0.22)
Model + LRCL (Ours, K = 5, S = 2,N = 4)	94.05% (+0.4)	98.72% (**+0.36**)	85.46% (+0.43)	90.03% (**+0.27**)
Model + LRCL (Ours, K = 5, S = 2, N = 8)	95.20% (**+1.55**)	98.68% (+0.30)	85.38% (+0.35)	89.98% (+0.22)
Model + LRCL (Ours, K = 5, S = 2, N = 16)	93.95% (+0.3)	98.66% (+0.28)	85.38% (+0.35)	89.90% (+0.14)

TABLE XI
PERFORMANCE OF IMPROVED LOCAL CONSTRAINT LOSS IN ACROSS DATASETS. K REPRESENTS THE SIZE OF THE BLOCK, S REPRESENTS THE STEP SIZE, AND N REPRESENTS THE NUMBER OF BLOCKS. THE RESULTS ARE TRAINED ON THE FF++ (C23) DATASET AND THE FF++ (C40) DATASET AND TESTED ON THE CELEB-DF, DFDC, AND WILDDEEPFAKE DATASETS. WE USE "/" TO SEPARATE THE EXPERIMENTAL RESULTS, WHERE THE LEFT SIDE REPRESENTS THE EXPERIMENTAL RESULTS OF THE C23 VERSION, AND THE RIGHT SIDE REPRESENTS THE EXPERIMENTAL RESULTS OF THE C40 VERSION.

Methods	Test Set (AUC)		
	Celeb-DF	DFDC	WildDeepfake
Model	71.77 / 66.22	69.46 / 63.60	70.16 / 67.07
Model + LRCL	71.28 (-0.49) / 66.62 (+0.40)	69.08 (-0.38) / 63.88 (+0.28)	70.73 (+0.57) / 67.27 (+0.20)
Model + LRCL (Ours, K = 3, S = 3, N = 1)	69.22 (-2.55) / 67.54 (+1.32)	69.34 (-0.12) / 63.71 (+0.11)	71.41 (+1.25) / 66.91 (-0.16)
Model + LRCL (Ours, K = 3, S = 3, N = 4)	69.42 (-2.35) / 66.13 (-0.09)	68.97 (-0.49) / 63.54 (-0.04)	70.96 (+0.8) / 67.24 (+0.17)
Model + LRCL (Ours, K = 3, S = 3, N = 8)	70.85 (-0.92) / 66.70 (+0.48)	69.59 (+0.13) / 64.09 (+0.49)	71.75 (+1.59) / 67.97 (+0.80)
Model + LRCL (Ours, K = 3, S = 3, N = 16)	69.63 (-2.14) / 66.48 (+0.26)	69.38 (-0.08) / 63.08 (-0.52)	70.68 (+0.52) / 67.26 (+0.19)
Model + LRCL (Ours, K = 5, S = 2, N = 1)	69.82 (-1.95) / 67.19 (+0.97)	69.15 (-0.31) / 64.00 (+0.40)	70.80 (+0.64) / 67.23 (+0.16)
Model + LRCL (Ours, K = 5, S = 2, N = 4)	69.84 (-1.93) / 66.53 (+0.31)	69.83 (**+0.37**) / 63.80 (+0.20)	71.63 (+1.47) / 67.06 (-0.01)
Model + LRCL (Ours, K = 5, S = 2, N = 8)	72.76 (**+0.99**) / 67.69 (**+1.47**)	69.57 (+0.11) / 64.33 (**+0.73**)	74.51 (**+4.35**) / 68.32 (**+1.25**)
Model + LRCL (Ours, K = 5, S = 2, N = 16)	69.55 (-2.22) / 66.54 (+0.32)	68.98 (-0.48) / 63.50 (-0.10)	70.66 (+0.5) / 67.34 (+0.27)

16. The specific experimental data can be found in Tables X and XI. The experimental data indicates that the local constraint losses with varying values in different scenarios can enhance the detection accuracy to some extent within the dataset. However, the differences are not very pronounced. Consequently, we conducted experiments to assess their cross-database performance. The results demonstrate that, under different step lengths and block sizes, an N value of 8 notably improves the generalization performance. To further analyze the reasons behind this improvement, we visualized different N values, and the results are presented in Fig. 7.

- When $N = 1$, the proposed model's focus using constrained loss is slightly different from the original. It merely emphasizes a noticeable artifact, potentially emphasizing the wrong area. For instance, in the NT subset, where artifacts mainly cluster around the lips, the model may incorrectly focus on the nose.
- When $N = 4$, the model concentrates on the more obvious artifacts, but its response to the subtle ones diminishes. For example, in the DF subset, the model may overlook subtle artifacts on the lower and upper left sides of the face.
- When $N = 8$, the model maintains its response to subtle artifacts while focusing on the obvious ones, potentially explaining the significant improvement in the generalization performance;
- When $N = 16$, the model struggles to detect images effectively. This is due to most fake images having fewer fake parts than real areas. Setting N to a large value causes the model to focus only on the real parts, contradicting the original intention behind the loss design.

As shown in Tables X and XI, when N = 16, the detection performance is lower than to when N = 8. Additionally, since N = 16 approaches the total number of blocks, we believe that further increasing the value of N may lead to a further decline in detection performance. Considering the trade-offs observed in our experiments, we ascertained that setting N = 8 balances between focusing on both the subtle and obvious artifacts. This selection ensures robust performance across different scenarios and datasets in our final experiments.

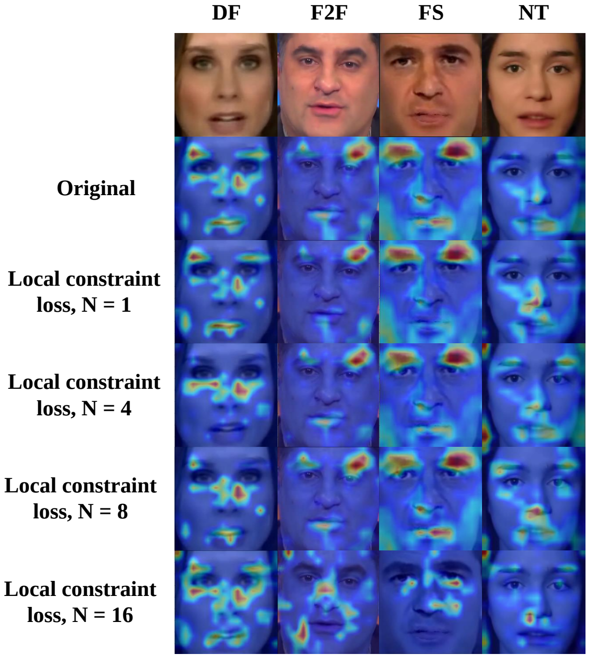

DF
F2F
FS
NT
Original
Local constraint loss, N = 1
Local constraint loss, N = 4
Local constraint loss, N = 8
Local constraint loss, N = 16

Fig. 7. Visualization results of local constraint loss under different values of N.

4) Comparison of the Computational Complexity of Existing Popular Networks: Finally, we analyzed the computational complexity (i.e., the FLOPs and parameter number) of various existing models. The specific experimental data are shown in Tab. XII.

TABLE XII
FLOPS AND PARAMETERS NUMBER FOR DIFFERENT METHODS.

Model	FLOPs	Param.
F3-Net [6]	18.1G	47.96M
GFFD [25]	13.81G	53.24M
M2TR [19]	4.6G	38M
Ours	2.13G	7.92M

According to the results, our method obtained the lowest throughput and parameter number compared to recent methods based on CNN and the Transformer. This highlights our approach's effectiveness and lightweight nature, making it a promising choice for applications where computational efficiency is crucial.

VI. Visualization and Analysis

We used the widely adopted Gradient Weighted Class Activation Mapping (Grad-CAM) technique [47] to visualize the feature maps, as depicted in Figure 8. We aim to illustrate our model's interpretability by presenting the heatmaps corresponding to the DFTM module, and the RGB and noise branches.

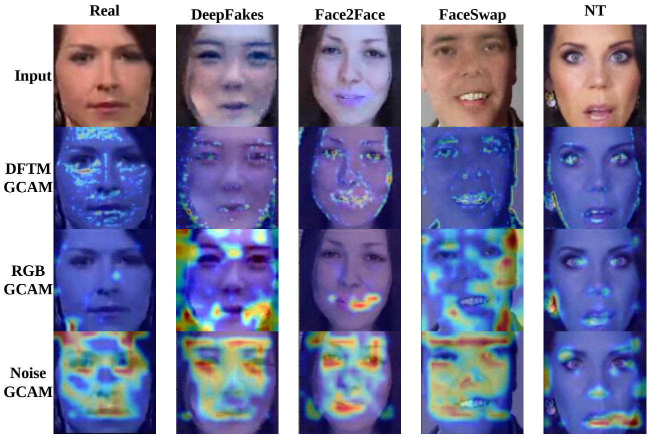


Fig. 8. Visualize the results. The second line is the heat map after being guided by the DFTM module; the third line is the heat map of the spatial flow; the fourth line is the heat map of the noise flow.

In the visualization second row, we observed that our DFTM module concentrates on high-frequency artifact information, particularly around the edges of false faces. Conversely, for genuine faces, the attention is more evenly distributed across the entire face. The RGB and noise streams play complementary roles in information processing. Guided by the DFTM module, the RGB stream focuses more on detailed artifact regions than the entire face. Meanwhile, the noise stream prioritizes global information, capturing noise inconsistencies from a broader perspective. The features extracted from these two streams complement each other, enabling the model to concentrate on various facial regions and explore forgery traces from multiple angles.

VII. Conclusions and Future Work

In this research, we presented a novel Two-Stream architecture designed specifically for face forgery detection. Considering the substantial artifact reduction inherent in recompressed data, our approach harnesses a multi-scale Transformer module to provide spatial guidance, facilitating the interpretation of frequency domain information and the detection of subtle artifact details. Moreover, we incorporated noise flow to complement spatial information, emphasizing inconsistencies in image noise patterns. Furthermore, a fusion module was introduced to enhance the feature refinement process. Our experimental evaluations, conducted across various public datasets, highlight our approach's superior performance compared to existing state-of-the-art methods. In future endeavors, we will focus on enhancing our models' computational efficiency, lightweight characteristics, and generalization performance.

References

[1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network." IEEE, 2018, pp. 1–7.

[2] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *ICASSP*, 2019, pp. 2307–2311.

[3] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *ICCV*, 2019, pp. 1–11.

[4] H. Mo, B. Chen, and W. Luo, "Fake faces identification via convolutional neural network," in *Proceedings of the 6th ACM workshop on information hiding and multimedia security*, 2018, pp. 43–47.

[5] Z. Guo, G. Yang, J. Chen, and X. Sun, "Fake face detection via adaptive manipulation traces extraction network," *Computer Vision and Image Understanding*, vol. 204, p. 103170, 2021.

[6] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *ECCV*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12357, 2020, pp. 86–103.

[7] G. Jia, M. Zheng, C. Hu, X. Ma, Y. Xu, L. Liu, Y. Deng, and R. He, "Inconsistency-aware wavelet dual-branch network for face forgery detection," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 3, pp. 308–319, 2021.

[8] X. Xu, W. Lv, W. Wang, Y. Zhang, and J. Chen, "Empowering semantic segmentation with selective frequency enhancement and attention mechanism for tampering detection," *IEEE Transactions on Artificial Intelligence*, 2023.

[9] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in *International conference on machine learning*. PMLR, 2020, pp. 3247–3258.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 5998–6008.

[11] J. Li, H. Xie, J. Li, Z. Wang, and Y. Zhang, "Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection," in *CVPR*, 2021, pp. 6458–6467.

[12] J. Li, H. Xie, L. Yu, and Y. Zhang, "Wavelet-enhanced weakly supervised local feature learning for face forgery detection," in *ACM*, J. Magalhães, A. D. Bimbo, S. Satoh, N. Sebe, X. Alameda-Pineda, Q. Jin, V. Oria, and L. Toni, Eds., 2022, pp. 1299–1308.

[13] J. J. Fridrich and J. Kodovský, "Rich models for steganalysis of digital images," *IEEE Trans. Inf. Forensics Secur.*, vol. 7, no. 3, pp. 868–882, 2012.

[14] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in *CVPR*, 2021, pp. 2185–2194.

[15] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, "On the detection of digital face manipulation," in *CVPR*, 2020, pp. 5780–5789.

[16] Z. Guo, L. Wang, W. Yang, G. Yang, and K. Li, "Ldfnet: Lightweight dynamic fusion network for face forgery detection by integrating local artifacts and global texture information," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[17] S. Chen, T. Yao, Y. Chen, S. Ding, J. Li, and R. Ji, "Local relation learning for face forgery detection," in *AAAI*, 2021, pp. 1081–1088.

[18] S. Chhabra, K. Thakral, S. Mittal, M. Vatsa, and R. Singh, "Low quality deepfake detection via unseen artifacts," *IEEE Transactions on Artificial Intelligence*, 2023.

[19] J. Wang, Z. Wu, W. Ouyang, X. Han, J. Chen, Y. Jiang, and S. Li, "M2TR: multi-modal multi-scale transformers for deepfake detection," in *ICMR*, V. Oria, M. L. Sapino, S. Satoh, B. Kerhervé, W. Cheng, I. Ide, and V. K. Singh, Eds., 2022, pp. 615–623.

[20] H. Chen, Y. Lin, B. Li, and S. Tan, "Learning features of intra-consistency and inter-diversity: Keys toward generalizable deepfake detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 3, pp. 1468–1480, 2022.

[21] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-stream neural networks for tampered face detection," in *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*. IEEE, 2017, pp. 1831–1839.

[22] Y. Nirkin, L. Wolf, Y. Keller, and T. Hassner, "Deepfake detection based on the discrepancy between the face and its context," *arXiv preprint arXiv:2008.12262*, 2020.

[23] J. Yang, S. Xiao, A. Li, W. Lu, X. Gao, and Y. Li, "Msta-net: Forgery detection by generating manipulation trace based on multi-scale self-texture attention," *IEEE transactions on circuits and systems for video technology*, vol. 32, no. 7, pp. 4854–4866, 2021.

[24] N. Park and S. Kim, "How do vision transformers work?" in *ICLR*. The organization, 2022.

[25] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *CVPR*, 2021, pp. 16 317–16 326.

[26] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *CVPR*, 2017, pp. 1800–1807.

[27] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *ICML*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97, 2019, pp. 6105–6114.

[28] C. Wang and W. Deng, "Representative forgery mining for fake face detection," in *CVPR*, 2021, pp. 14 923–14 932.

[29] Z. Liu, X. Qi, and P. H. S. Torr, "Global texture enhancement for fake face detection in the wild," in *CVPR*, 2020, pp. 8057–8066.

[30] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. 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*, 2021, pp. 772–781.

[31] C. Miao, Z. Tan, Q. Chu, N. Yu, and G. Guo, "Hierarchical frequency-assisted interactive networks for face manipulation detection," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 3008–3021, 2022.

[32] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, "Wilddeepfake: A challenging real-world dataset for deepfake detection," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 2382–2390.

[33] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2185–2194.

[34] Z. Guo, G. Yang, D. Zhang, and M. Xia, "Rethinking gradient operator for exposing ai-enabled face forgeries," *Expert Systems with Applications*, vol. 215, p. 119361, 2023.

[35] C. Miao, Z. Tan, Q. Chu, H. Liu, H. Hu, and N. Yu, "F 2 trans: High-frequency fine-grained transformer for face forgery detection," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1039–1051, 2023.

[36] Z. Guo, Z. Jia, L. Wang, D. Wang, G. Yang, and N. Kasabov, "Constructing new backbone networks via space-frequency interactive convolution for deepfake detection," *IEEE Transactions on Information Forensics and Security*, 2023.

[37] D. Zhang, J. Chen, X. Liao, F. Li, J. Chen, and G. Yang, "Face forgery detection via multi-feature fusion and local enhancement," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[38] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *CVPR*, 2020, pp. 3204–3213.

[39] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge (dfdc) dataset," *arXiv preprint arXiv:2006.07397*, 2020.

[40] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Faceshifter: Towards high fidelity and occlusion aware face swapping," *arXiv preprint arXiv:1912.13457*, 2019.

[41] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, 2009.

[42] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Z. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *NeurIPS*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 8024–8035.

[43] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam," *CoRR*, vol. abs/1711.05101, 2017.

[44] Y. Guo, C. Zhen, and P. Yan, "Controllable guide-space for generalizable face forgery detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 818–20 827.

[45] Y. Xu, K. Raja, L. Verdoliva, and M. Pedersen, "Learning pairwise interaction for generalizable deepfake detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 672–682.

[46] Z. Yan, Y. Zhang, Y. Fan, and B. Wu, "Ucf: Uncovering common features for generalizable deepfake detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 412–22 423.

[47] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017, pp. 618–626.

[48] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.