electronics

MDPI

Article

Refining Localized Attention Features with Multi-Scale Relationships for Enhanced Deepfake Detection in Spatial-Frequency Domain

Yuan Gao [1,2], Yu Zhang [1,3], Ping Zeng [1,3,*] and Yingjie Ma [1]

[1] Department of Electronics and Communications Engineering, Beijing Electronic Science and Technology Institute, Beijing 100070, China; gy@besti.edu.cn (Y.G.); zhangyu21@stu.xidian.edu.cn (Y.Z.); myj@besti.edu.cn (Y.M.)
[2] State Information Center, Beijing 100045, China
[3] School of Telecommunications Engineering, Xidian University, Xi'an 710071, China
* Correspondence: zp@besti.edu.cn

Abstract: The rapid advancement of deep learning and large-scale AI models has simplified the creation and manipulation of deepfake technologies, which generate, edit, and replace faces in images and videos. This gradual ease of use has turned the malicious application of forged faces into a significant threat, complicating the task of deepfake detection. Despite the notable success of current deepfake detection methods, which predominantly employ data-driven CNN classification models, these methods exhibit limited generalization capabilities and insufficient robustness against novel data unseen during training. To tackle these challenges, this paper introduces a novel detection framework, ReLAF-Net. This framework employs a restricted self-attention mechanism that applies self-attention to deep CNN features flexibly, facilitating the learning of local relationships and inter-regional dependencies at both fine-grained and global levels. This attention mechanism has a modular design that can be seamlessly integrated into CNN networks to improve overall detection performance. Additionally, we propose an adaptive local frequency feature extraction algorithm that decomposes RGB images into fine-grained frequency domains in a data-driven manner, effectively isolating fake indicators in the frequency space. Moreover, an attention-based channel fusion strategy is developed to amalgamate RGB and frequency information, achieving a comprehensive facial representation. Tested on the high-quality version of the FaceForensics++ dataset, our method attained a detection accuracy of 97.92%, outperforming other approaches. Cross-dataset validation on Celeb-DF, DFDC, and DFD confirms the robust generalizability, offering a new solution for detecting high-quality deepfake videos.

Keywords: deepfake detection; deep learning; deepfake; local relationships; self-attention; fine-grained frequency features

Citation: Gao, Y.; Zhang, Y.; Zeng, P.; Ma, Y. Refining Localized Attention Features with Multi-Scale Relationships for Enhanced Deepfake Detection in Spatial-Frequency Domain. *Electronics* **2024**, *13*, 1749. https://doi.org/10.3390/electronics13091749

Academic Editor: Duc Thanh Nguyen

Received: 14 April 2024
Revised: 28 April 2024
Accepted: 30 April 2024
Published: 1 May 2024

1. Introduction

Face forgery technology has evolved into deepfake, which leverages Generative Adversarial Networks (GANs) [1] to create highly realistic faces by learning from extensive sample sets, a significant advancement over basic stitching and synthesis techniques. While these technologies offer vast potential for entertainment and media creation, they are frequently misused to produce counterfeit faces, including for mocking political figures and celebrities, manipulating public opinion, and committing fraud. This misuse has intensified public concerns regarding personal image theft, identity forgery, and the spread of misinformation on social platforms, leading to serious trust and security issues. Consequently, the development of effective deepfake detection technologies is imperative.

Currently, the primary methods [2–4] treat deepfake detection as a binary classification task, where neural networks learn specific features of forged videos and subsequently

classify the videos as either real or fake. However, this approach is susceptible to overfitting, leading to poor generalization performance and limited interpretability. To identify critical factors in CNNs judgment on forgeries, Chai et al. [5] have highlighted the essential features for detecting forgery by truncating CNN features and applying patch operations to predict each local feature. This research has paved the way for a new method in deepfake detection focusing on local features. Furthermore, Dong et al. [6] have demonstrated that the generalization challenges of binary classifiers in deepfake detection stem from the inadvertent learning of identity representations in images, a problem known as implicit identity leakage. A straightforward and effective solution involves performing forgery detection based on local features to minimize the reliance on global identity information, thus enhancing the model's generalization capabilities.

Through the analysis of the FaceForensics++ (FF++) [3] dataset, we have discovered that manipulated facial images display two distinct characteristics. Firstly, forged faces retain most of the original face's regions; for instance, technologies like Deepfakes [7] and FaceSwap [8] modify facial identification information while preserving the background, which remains unaltered. In contrast, Face2Face [9] and NeuralTextures [10] primarily alter facial expressions or lip movements without changing other areas. Secondly, the areas manipulated are predominantly local rather than global. The samples in the FF++ dataset are shown in Figure 1. These observations lead us to conclude that the extraction of local features and the understanding of local relationships are crucial and effective strategies for detecting facial manipulations.

Figure 1. Two sets of facial images randomly selected from the FaceForensics++ [3] dataset, where the two faces in each set were manipulated using Deepfakes, FaceSwap, Face2Face, and NeuralTextures, respectively.

Several researchers have explored local relationships to identify patterns conducive to forgery detection. Chen et al. [11] developed a Multi-scale Patch Similarity Module to assess the similarity between local features. Yang et al. [12] advanced masked relationship learning, utilizing spatio-temporal attention to analyze features across multiple facial regions and disseminate relationship information, thereby detecting irregularities from a global perspective. These approaches, however, rely on manually designed similarity patterns, which may not fully exploit local features and could render classification models biased toward specific types of forgery traces, limiting their generalization capabilities. Nevertheless, the advent of Transformers [13] and Vision Transformer (ViT) [14] has underscored their efficacy in various applications. Utilizing the core self-attention module of Transformers to model relationships between local features holds significant promise.

Compared to conventional CNNs, Transformers offer enhanced flexibility in modeling long-range dependencies in visual tasks, introduce minimal inductive bias, and are superior in extracting relational representations from local facial features to discern authenticity. Miao et al. [15] employed a bag-of-features approach to encode patch relationships, but relying solely on the Transformer encoder proved inadequate for comprehensively learning the subtle artifacts present on forged faces.

Recent studies have leveraged frequency clues to enhance the robustness of detection models against external disturbances. Frank et al. [16] and Liu et al. [17] have noted that most facial manipulation methods employ GANs, where the upsampling processes may lead to abnormal frequency statistical properties in forged faces, thus enhancing the robustness of forgery detection based on frequency domain features. However, current methods for extracting frequency features are relatively crude, generally relying on Discrete Cosine Transform (DCT) and manually designed filters. These methods convert the processed frequency domain features back to the RGB domain before integrating them into CNN models. Therefore, a fine-grained approach to frequency feature extraction could facilitate the identification of subtle facial forgery discriminative features, thereby enabling the network to better distinguish between authentic and manipulated areas.

To address the issues discussed earlier, this paper introduces a detection method that utilizes multi-scale local relationships in both the spatial and frequency domains, making significant contributions in several key areas:

- An adaptive local frequency feature extraction algorithm is introduced. It performs fine-grained frequency domain decomposition of RGB images, effectively isolating forgery traces in the frequency space and providing a robust representation of local feature relationships. This sets a solid foundation for the subsequent restricted local attention module.
- A novel dual-channel feature fusion module is proposed, which adaptively integrates RGB and frequency domain information, ensuring a comprehensive representation of facial features. This enriches the model's input with both spatial and spectral data, thereby enhancing detection accuracy.
- The paper innovatively improves upon self-attention mechanisms through the development of the Restricted Local Attention Module (ReLAM). ReLAM meticulously constrains the scope of self-attention, facilitating the learning of fine-grained local relationships and global feature dependencies in two distinct phases.

2. Related Work

Through experimental observation and prior knowledge, several heuristic methods have been developed to tackle the challenge of generalization in forgery detection. Due to the heterogeneity between authentic and forged faces, some studies employ mixed boundary artifacts for detection. Li et al. [18] introduced Face X-ray to determine whether an image is composed of parts from different sources, exposing the mixed boundaries of forged images. Zhao et al. [19] proposed a pair-wise self-consistency learning method, exploiting inconsistencies in source features of forged images for detection. Similarly, anomalies in the frequency domain of forged images are utilized for this purpose. Qian et al. [20] devised two innovative methods for frequency feature extraction: Frequency-Aware Decomposition and Local Frequency Statistics. Li et al. [21] developed an adaptive frequency feature generation module, while Gu et al. [22] proposed a progressive enhancement learning framework that integrates RGB information with fine-grained frequency clues. Additionally, Sun et al. [23] found that most forgery clues are located in high-information areas, quantifiable through classical information maximization theory, and introduced self-information measurement to improve feature representation for forgery detection. Some works [24–27] differentiate real from forged faces by analyzing the consistency of identity features inherent in faces, focusing on high-level identity features. While these heuristic methods enhance the detection of forged faces by incorporating specific domain knowledge, they lack unified theoretical support and may introduce biases into the models.

Recent research has focused on attention mechanisms. Dang et al. [28] utilized these mechanisms to analyze CNN feature maps, pinpointing image regions that impact CNN decisions via learned attention maps. Zhao et al. [29] approached forgery detection as a fine-grained classification issue, developing a multi-attention detection network. Fei et al. [30] introduced a weakly supervised second-order local anomaly learning model that leverages deep feature maps to identify anomalous features in local areas. Additionally, methods that focus on local relationships [31–33] have gained significant attention, differentiating between original and forgery-related features by modeling local feature interactions using various strategies. However, most of these studies require additional pixel-level fake position annotations for supervision or rely solely on a single strategy for learning local relationships, which hinders the full exploitation of local interactions and restricts potential enhancements.

Data augmentation has been employed to enhance the generalization capabilities of CNN models. Wang et al. [34] introduced an attention-based data augmentation strategy called RFM, designed to refine training data during the learning process. Zhu et al. [35] implemented a 3D decomposition technique to segregate face images into five distinct components, amplifying the primary indicators for forgery detection. However, the use of additional generative networks increases the computational load, and the noise introduced must be meticulously managed, particularly in cases of extreme forgery. Additionally, Wang et al. [36] detected artificially synthesized fake faces by analyzing neuron activity, while Luo et al. [37] utilized high-frequency features to identify forgeries. Although these methods significantly improve deepfake detection, the challenge of identifying unknown forgeries persists.

The aforementioned methods are primarily employed to identify forgery traces within video frames. However, since a forged video is composed of a series of tampered single-frame images, it inevitably leaves traces of forgery between frames. Sun et al. [38] utilized 68 facial keypoints in each frame and calculated the temporal features between frames based on optical flow algorithm, subsequently employing an RNN network to generate video-level detection results. Zheng et al. [39] leveraged temporal coherence for face forgery detection, introducing a Full Temporal Convolution Network and Temporal Transformer to identify temporal inconsistencies in forged videos. Hu et al. [40] developed a detection method based on video frame inference, reconceptualizing the deepfake detection into a face frame sequence inference task, thus providing a new perspective for deepfake detection. Further, several studies have detected temporal feature anomalies by monitoring physiological signals. Jung et al. [41] determine video authenticity by examining eye blink frequency and duration, and Qi et al. [42] assess forgery through observed heartbeat rhythms. Haliassos et al. [43] introduced the LipForensics method, which focuses on capturing detailed representations of mouth movements during speech. Although inter-frame detection methods are theoretically more effective, their practical performance is often similar similar to frame-level methods. Consequently, video sequence-based detection methods warrant additional research.

3. Methods

The overall model architecture, depicted in Figure 2, comprises three primary modules: the Adaptive Local Frequency Extraction Module (ALFEM), the Dual-Channel Feature Fusion Module (DCFFM), and the Restricted Local Attention Module (ReLAM). Specifically, ALFEM captures fine-grained frequency domain features from input facial RGB images. DCFFM integrates deep CNN features from both spatial and frequency domains to ensure a comprehensive representation of facial information. ReLAM analyzes multi-scale local relationships within the fused deep features, specifically targeting notably discordant local features. Detailed descriptions of each module follow.

3.1. Adaptive Local Frequency Extraction Module

Most existing deepfake detection methods that utilize frequency domain features do so in a coarse-grained manner. Inspired by the local frequency statistics method in F^3-Net [20], we have developed the Adaptive Local Frequency Extraction Module (ALFEM). This module performs fine-grained decomposition of RGB images, effectively isolating forged traces in the frequency space. As illustrated in Figure 2, ALFEM initially processes the input facial image by segmenting it into patches. For each patch, the Discrete Cosine Transform (DCT) extracts spectral coefficients, which are subsequently reconstituted into the original space after being filtered. Depthwise separable convolutions are then employed to directly extract fine-grained local frequency features within the frequency domain. This innovative design adaptively investigates fine-grained frequency domain representations within identical frequency bands and elucidates the importance of various frequency band features in detecting forgery clues.

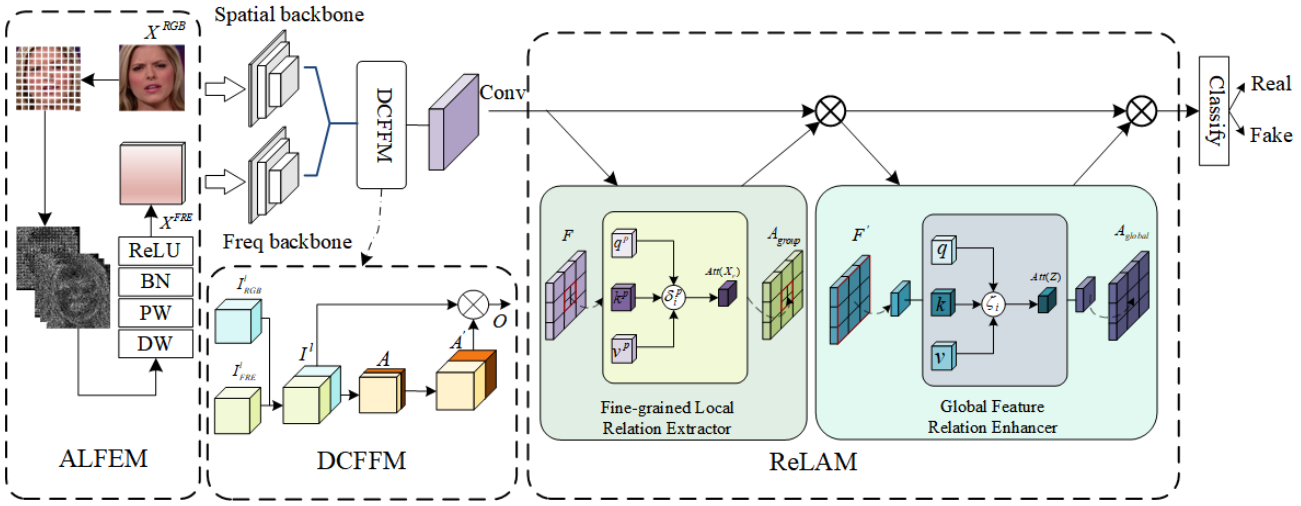


Figure 2. The overall architecture of ReLAF-Net.

Without loss of generality, let $X^{RGB} \in \mathbb{R}^{3\times H\times W}$ represent the input RGB image, where H and W denote the height and width, respectively. First, X^{RGB} is processed with a sliding window slice to obtain a set of local patchs of size $S \times S$, where S is the size of the sliding window, and $P^{rgb}_{(i,j)} \in \mathbb{R}^{3\times S\times S}$ represents the patch at position (i, j). Then, each $P^{rgb}_{(i,j)}$ undergoes a Discrete Cosine Transform to obtain a spectral representation $P^{fre}_{(i,j)} \in \mathbb{R}^{1\times S\times S}$. Subsequently, it passes through n filters of different frequency bands, and the filtered spectral maps containing only specified frequency components are stacked together. The specific implementation can be described as follows:

$$P^{FRE}_{(i,j)} = \text{Concat}(f^k * P^{fre}_{(i,j)}) = \text{Concat}(f^k * \mathcal{DCT}(P^{rgb}_{(i,j)})), \quad k = 1, \ldots, n, \tag{1}$$

where, $P^{FRE}_{(i,j)} \in \mathbb{R}^{n\times S\times S}$ is the local spectral representation at position (i, j), Concat$(\cdot)$ denotes concatenating feature maps along the channel direction. f^k represents the kth filter in the filter group, implemented based on a 0–1 mask map, with the ratio of 0–1 regions adjusted to control the range of frequency bands to be filtered. Figure 3 shows the composition of the mask map, which is the same size as the patch size. Assuming the top-left coordinate of the input image is $(0, 0)$, then the kth mask map $mask_k$ has value 1 within the area formed by coordinates $(0, (k-1)H/n)$, $(0, kH/n)$, $((k-1)W/n, 0)$, $(kW/n, 0)$, and 0 elsewhere.

Subsequently, $P^{FRE}_{(i,j)}$ is reassembled according to its original spatial position to generate a frequency domain feature representation $B^{FRE} \in \mathbb{R}^{n\times(S\times H')\times(S\times W')}$, where H' and W' are the vertical and horizontal steps of the sliding window. Finally, local frequency domain

information is adaptively extracted through depthwise separable convolution to form a fine-grained frequency representation $X^{FRE} \in \mathbb{R}^{C' \times H' \times W'}$, as follows:

$$X^{FRE} = \text{ReLU}(\text{BN}(DW(B^{FRE}))), \tag{2}$$

where, ReLU is the activation function, BN is batch normalization. DW denotes depthwise separable convolution, with the kernel size of the Depthwise operation set to $S \times S$, stride to S, and the number of groups to n. The kernel size of the Pointwise operation is 1×1, used for adaptively extracting features of different frequency bands.

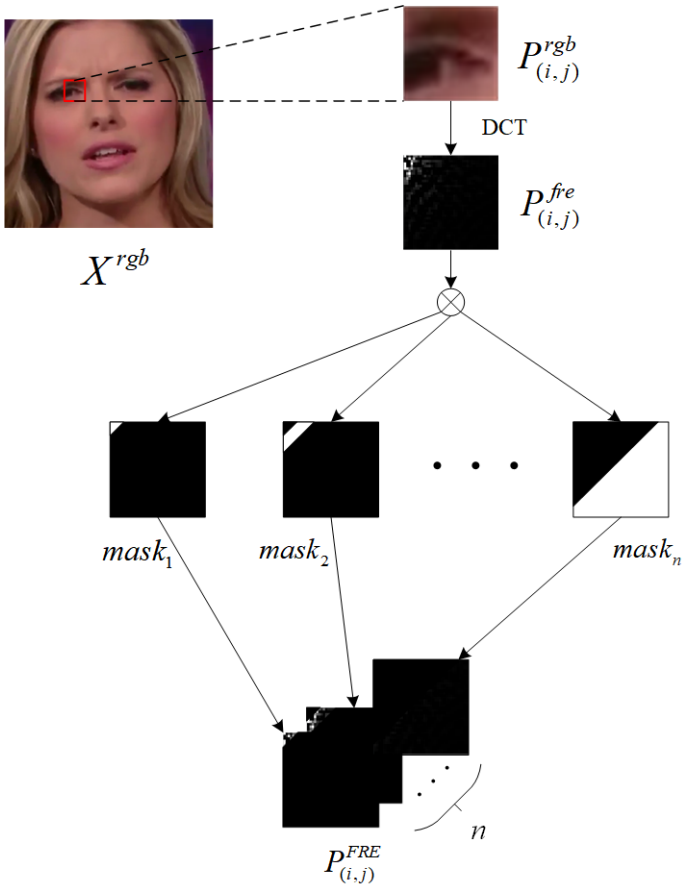


Figure 3. The process of computing fine-grained frequency features from RGB images.

ALFEM decomposes the original input and reconstructs it into fine-grained frequency-aware data. This process effectively exposes local anomalies in forged images across various frequency bands while preserving spatial relationships to accommodate shift invariance. Consequently, ALFEM establishes a solid foundation for extracting detailed features from both authentic and manipulated facial regions.

3.2. Dual-Channel Feature Fusion Module

To effectively capture the anomalous textures and subtle manipulations in forged faces, we have introduced a Dual-Channel Feature Fusion Module that seamlessly integrates RGB and frequency information, ensuring a comprehensive representation of facial data. The structure of the DCFFM is depicted in Figure 2.

Original RGB images X^{RGB} and the fine-grained frequency features X^{FRE} produced by the ALFEM module are processed through a parallel dual-channel backbone network to extract deep features. Each channel independently processes its respective feature maps until the l-th layer, where we obtain $I^l_{RGB} \in \mathbb{R}^{C \times H \times W}$ and $I^l_{FRE} \in \mathbb{R}^{C \times H \times W}$, representing

the feature maps of the RGB and frequency branches, respectively. The objective is to integrate these dual-branch features.

First, these two features are concatenated along the channel dimension to obtain the fused feature map $I^l \in \mathbb{R}^{2C\times H\times W}$, expressed as

$$I^l = [I^l_{RGB}; I^l_{FRE}], \tag{3}$$

The fused feature map I^l goes through a series of operations including a 1×1 convolution layer, Batch Normalization, and ReLU non-linear mapping, followed by a 3×3 convolution layer and a sigmoid function to derive the attention map $A \in \mathbb{R}^{2\times H\times W}$. The complete process can be represented as

$$V = \text{ReLU}\{\text{BN}\{\text{Conv}_{1\times 1}(I^l)\}\}, \tag{4}$$

$$A = \text{sigmoid}(\text{Conv}_{3\times 3}(V)), \tag{5}$$

Here, A contains two channels of attention weights, specifically represented as $A_1 \in \mathbb{R}^{1\times H\times W}$ and $A_2 \in \mathbb{R}^{1\times H\times W}$. Then, repeat each channel C times and concatenate them according to the original channel order to form a new feature tensor $A' \in \mathbb{R}^{2C\times H\times W}$. Finally, the feature I^l and the attention vector A' are fused through a weighted combination.

$$A' = [Repeat(A_1); Repeat(A_2)], \tag{6}$$

$$O = I^l + I^l \odot A', \tag{7}$$

where, $\odot$ represents element-wise multiplication, and the output $O \in \mathbb{R}^{2C\times H\times W}$ represents the robustly fused features derived from both RGB and frequency data, followed by remaining processing via a single channel.

DCFFM can adaptively learn the dependency between RGB and frequency information. By integrating these two distinct types of data, the model develops a comprehensive understanding, which enhances feature representation for downstream tasks and ensures robust facial forgery detection.

3.3. Restricted Local Attention Module

To understand local relationships of forgery patterns at different scales, we have developed an innovative attention mechanism named Restricted Self-Attention. This mechanism confines the scope of self-attention to deep CNN features, thereby enabling the learning of multi-scale local relationships and improving the detection of subtle forgery traces.

As depicted in Figure 2, ReLAM comprises two components: (1) Fine-grained Local Relation Extractor, which identifies subtle manipulation traces by extracting fine-grained feature relationships within local patches; and (2) Global Feature Relation Enhancer, which learns inter-regional relationships in groups to better extract deep dependency relationships. By modularizing restricted self-attention as ReLAM, we facilitate its seamless integration into CNN networks, significantly enhancing detection performance and generalization capacity.

3.3.1. Fine-Grained Local Relation Extractor

As shown in Figure 4a, given an input feature map $F \in \mathbb{R}^{C\times H\times W}$, where C is the number of channels and the spatial dimensions are $H \times W$, the input feature is divided into multiple local regions in the spatial dimensions, with each local region being of size $H' \times W'$. This results in a total of $H/H' \times W/W'$ local patches. The Fine-grained Local Relation Extractor limits the computation scope of self-attention within each local area $X_p \in \mathbb{R}^{C\times H'\times W'}$.

First, compute the input query vector q_p, key vector k_p, and value vector v_p for self-attention as follows:

$$q_p = X_p W_p^Q, \quad k_p = X_p W_p^K, \quad v_p = X_p W_p^V, \tag{8}$$

Next, apply multi-head attention within the local region X_p to learn fine-grained local relationships. The calculation process is as follows:

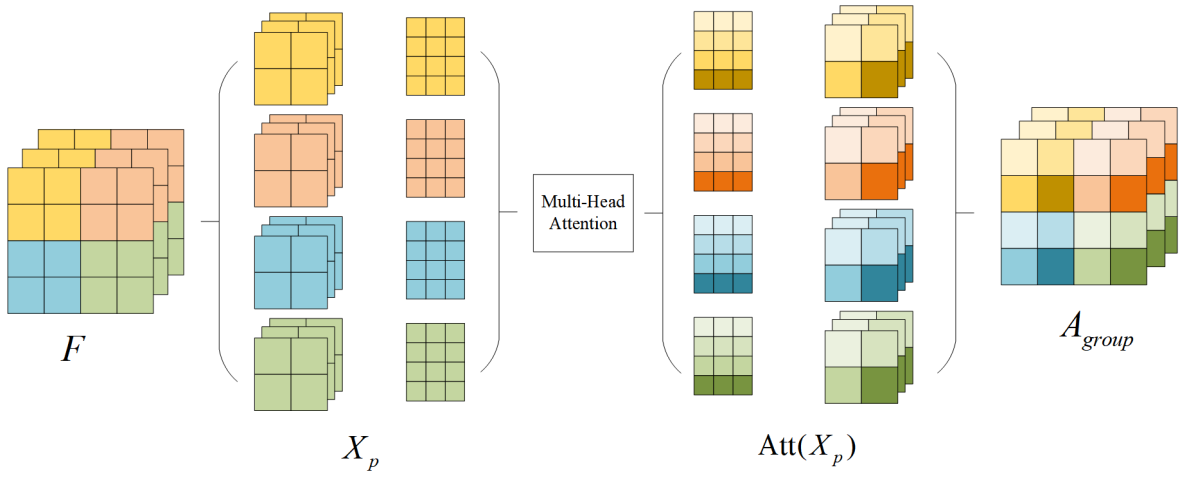
$$\mathrm{Att}(X_p) = \mathrm{MultiHead}(q_p, k_p, v_p), \tag{9}$$

For each local area $X_p^{(i,j)}$, the calculated $\mathrm{Att}(X_p^{(i,j)})$ is reassembled according to its position in the original feature F, thus obtaining an attention map A_{group} that represents fine-grained local relationships, expressed as follows:

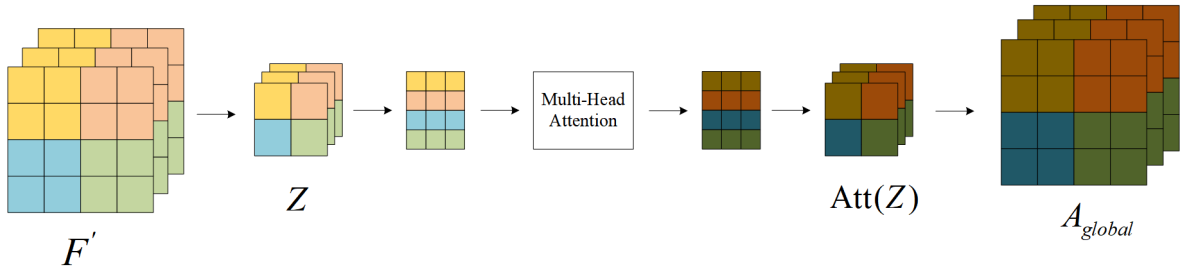
$$A^{group}_{(i\times W', j\times H')\to((i+1)\times W', (j+1)\times H')} = \mathrm{Att}(X_p^{(i,j)}), \tag{10}$$

Finally, A_{group} and the input feature F are connected via a residual connection, and the output of the extractor $F' \in \mathbb{R}^{C\times H\times W}$ is as shown below:

$$F' = F + F \odot A_{group}, \tag{11}$$



(a) Fine-grained Local Relation Extractor



(b) Global Feature Relation Enhancer

Figure 4. The dimensional transformation process of feature embedding in ReLAM. (**a**) is the process of Fine-grained Local Relation Extractor. (**b**) is the process of Global Feature Relation Enhancer.

3.3.2. Global Feature Relation Enhancer

For the output F' of the Fine-grained Local Relation Extractor, first use two layers of depthwise separable convolutions for global subsampling. The sparse connections of the depthwise separable convolutions allow for effectively learning representations for each group, as shown below:

$$Z = \sigma(DW_1(\sigma(DW_0(F')))), \tag{12}$$

Here, DW_i represents the ith layer of the depthwise separable convolution, $Z \in \mathbb{R}^{C\times\hat{H}\times\hat{W}}$, where $\hat{H} = \frac{H}{H'}$ and $\hat{W} = \frac{W}{W'}$, respectively, represent the number of patches in the vertical and horizontal directions.

Next, apply multi-head attention to the feature Z, thereby learning dependencies between regions at the global feature level, obtaining a global attention map $\mathrm{Att}(Z) \in \mathbb{R}^{C\times\hat{H}\times\hat{W}}$, as shown in Figure 4b.

$$A^{global}_{(i\times W',j\times H')\to((i+1)\times W',(j+1)\times H')} = \widehat{\mathrm{Att}}(Z^{(i,j)}) = \mathrm{Fill}(\mathrm{Att}(Z)_{i,j}), \tag{13}$$

The feature of $\mathrm{Att}(Z)$ at the spatial position (i, j) can be represented as $Att(Z)_{i,j} \in \mathbb{R}^{C\times1\times1}$. By filling $Att(Z)_{i,j}$ to transform it into $\widehat{Att}(Z)_{i,j} \in \mathbb{R}^{C\times H'\times W'}$. Then, combine all $\widehat{Att}(Z)_{i,j}$ according to their original positions to obtain the attention map $A_{global} \in \mathbb{R}^{C\times H\times W}$, as described in Equation (13). Finally, A_{glabal} and F' are connected via a residual connection, as follows:

$$F'' = F' + F' \odot A_{global}, \tag{14}$$

Additionally, before both the Fine-grained Local Relation Extractor and the Global Feature Relation Enhancer, convolution-based positional encoding is introduced, as follows:

$$\mathrm{PEG}(x) = x + \mathrm{Conv}(x), \tag{15}$$

The Restricted Local Attention Module measures the relevance of each segment of the input information to others, enhancing relevant areas and suppressing less pertinent ones. Consequently, this module proficiently learns the dependencies between subtle facial features and can be effectively used for distinguishing between genuine and fake faces, thereby boosting the model's discriminative power and generalization capabilities.

3.4. Loss Function

This method uses the cross-entropy loss function, which is expressed as follows:

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

where N represents the number of samples, y_i is the actual label of the ith sample, and $\hat{y}_i$ is the corresponding sample's predicted value.

4. Experiments and Results Analysis

This section describes the experimental setup and provides the results of the comparison and ablation experiments, which are analyzed and summarized.

4.1. Experimental Setup

4.1.1. Datasets

With the proliferation of deepfake content, related research has also emerged, and more and more public datasets have been released, aimed at advancing the detection and defense research against deepfakes. We evaluate the proposed method on four widely used public datasets, including FF++ [3], Celeb-DF [44], DFD [45], and DFDC [46], where the first two datasets are used for training and evaluation, and the latter two are used only for cross-dataset evaluation.

- The FF++ dataset contains 1000 real videos and 4000 fake videos created using four different face tampering techniques, including DeepFakes, FaceSwap, Face2Face, and Neural Textures. Additionally, for all videos, three versions of compression treatment using the H.264 codec are provided: C0 (raw), C23 (HQ), and C40 (LQ). The diversity of the FF++ dataset can evaluate the model's generalization ability and robustness to unknown forgery methods and videos of different compression levels, allowing for a

comprehensive assessment of the performance and robustness of deepfake detection models. Therefore, the FF++ dataset plays an important role in the field of deepfake detection, providing strong support for the development of related fields.

- The Celeb-DF dataset contains 890 real videos and 5639 fake videos, where the fake videos were generated using an improved DeepFake algorithm, greatly reducing visual artifacts and possessing higher visual quality.
- The DFD dataset contains 363 original videos shot in 16 different scenes by 28 actors and 3068 fake videos generated using deep learning methods, with the specific forgery method not disclosed.
- The DFDC dataset contains 23,654 real videos and 104,500 fake videos. Due to the variety and unknown nature of the forgery methods, and the presence of various complex scenes, it is very challenging for existing forgery detection.

The original videos are stored in MP4 format. For frame-based training, we adhere to established protocols by utilizing OpenCV to extract the first 300 frames from each video, from which we randomly select 50 frames for training and testing. Subsequently, the Dlib library is employed to detect and align faces within these frames. All facial images are then cropped and normalized to 299 × 299 pixels and saved in PNG format. For the FF++ dataset, following the settings in reference [3], we divided the dataset into training, validation, and testing sets in a ratio of 720:140:140. The division of the training and test sets for the Celeb-DF dataset followed the official guidelines. The DFD and DFDC datasets are used exclusively for cross-dataset evaluation experiments. Due to the imbalance in sample distribution, in the DFD dataset, we selected all 363 real videos and one-tenth of the fake videos. Considering the large volume of the DFDC dataset, only 1000 videos were selected to construct the testing set. Table 1 shows the statistical information of the frames collected from each dataset. Additionally, during the training process using FF++ and Celeb-DF, a series of data augmentation techniques were applied to the positive samples due to the imbalance issue between positive and negative samples. Specifically, horizontal flipping, random cropping, color transformation, and combinations of these methods were used to increase the number of positive samples, thereby achieving sample balance.

Table 1. Frame number statistics of the datasets used for training, validation, and testing.

Datasets	Label	Train (Frame)	Valid (Frame)	Test (Frame)
FF++	Real	36k	7k	7k
	Fake	144k	28k	28k
Celeb-DF	Real	35.6k	-	8.9k
	Fake	264.95k	-	17k
DFD	Real	-	-	18.15k
	Fake	-	-	15.3k
DFDC	Real	-	-	25k
	Fake	-	-	25k

4.1.2. Evaluation Metrics

This paper proposes a frame-level detection method for videos, hence the evaluation is conducted at the image level, using accuracy (ACC), the receiver operating characteristic curve (ROC), and the area under the curve (AUC) as the model performance evaluation metrics. The experimental results of other methods are directly cited for comparison.

4.1.3. Implementation Details

The dual-branch backbone network employs the Xception [47] pre-trained on ImageNet. Xception has 14 residual convolutional blocks, with the first convolutional layer in the frequency domain branch replaced by ALFEM, DCFFM inserted after the 10th block, followed by a 1 × 1 convolution and ReLAM, and subsequent Xception structure serving as the classification layer. The input RGB image shape is 3 × 299 × 299, the sliding window

for DCT transform is 10×10, and the number of filters in the filter group is 6. The output dimension of the ALFEM module is $32 \times 149 \times 149$, and the output dimension of DCFFM is $1024 \times 10 \times 10$. Moreover, restricted self-attention is implemented in the form of multi-head attention, with the restricted window size being 2×2, the number of heads being 8, and the embedding dimension per attention head being 128.

All experiments were conducted on an Ubuntu 20.04 system using the Pytorch and NVIDIA RTX 3090 24 GB. During the network training process, the Adam optimizer was used, where β_1 and β_2 are set to 0.9 and 0.999, respectively, with an initial learning rate of 0.0002, and weight decay of 1×10^{-8}. The batch size is 32, the number of epochs is set to 50, and a StepLR scheduler is adopted, halving the learning rate after every 5 epochs. Additionally, an early stopping strategy is employed to store the best model weights, to prevent model overfitting.

4.2. Quantitative Results

4.2.1. In-Dataset Evaluations

Evaluations were conducted within the datasets using two qualities (HQ and LQ) of FF++ and the Celeb-DF dataset. As shown in Tables 2 and 3, our method demonstrated superior performance on both datasets. Specifically, on the FF++ dataset, compared to methods based on local relations, such as LRL and MRL, our method increased the average AUC by 1.26% and 1.37% for the two different compression ratios, respectively. Moreover, the proposed scheme also showed significant improvements over methods based on frequency features, such as F^3-Net, SPSL, and Freq-SCL, as well as the attention-based method MAT. Experimental results on the Celeb-DF dataset were also superior to all comparison schemes. Therefore, the in-dataset experimental results demonstrate the effectiveness of our proposed detection method.

Table 2. Quantitative results on FaceForensics++ dataset with different qualities (HQ and LQ).

Methods	FF++ (LQ)		FF++ (HQ)	
	ACC (%)	AUC (%)	ACC (%)	AUC (%)
Xception [3]	86.86	89.30	95.73	96.30
Face X-ray [18]	-	61.10	-	87.40
F^3-Net [20]	90.43	93.30	97.52	98.10
SPSL [17]	81.57	82.82	91.50	95.32
Freq-SCL [21]	89.00	92.40	96.69	99.30
MAT [29]	88.69	90.40	97.60	99.29
SIA [23]	90.23	93.45	97.64	99.35
RECCE [25]	91.03	95.02	97.06	99.32
BOF [15]	87.86	91.61	96.57	99.36
LRL [11]	91.47	95.21	97.59	99.46
MRL [12]	91.81	96.18	93.82	98.27
OUR	93.95 *	97.48 *	97.92 *	99.72 *

The highest results are highlighted with *.

Table 3. Quantitative results on Celeb-DF dataset.

Methods	ACC (%)	AUC (%)
Xception [3]	97.90	99.73
F^3-Net [20]	95.95	98.93
MAT [29]	97.92	99.94
SIA [23]	98.48	99.64
RECCE [25]	98.59	99.94
MRL [12]	-	99.96
SFDG [48]	99.22	99.96
OUR	99.60 *	99.98 *

The highest results are highlighted with *.

4.2.2. Cross-Manipulation Evaluations

The FF++ dataset explicitly divides four different forgery methods into subsets, which enables the study of generalization to unknown forgery methods. Based on the experimental setup of Freq-SCL [21], we conducted cross-manipulation experiments on the FF++ (HQ) dataset, with AUC (%) as the evaluation metric.

As shown in Table 4, our method achieves higher average AUC on the four forgery subsets when trained on datasets containing only one type of forgery method, compared to the comparative methods. Also, in most individual comparisons, it performed better than the comparative methods. Notably, models trained on DF and NT and tested on FS and F2F showed results below RECCE. We speculate that the possible reason is that FS and F2F are forgeries based on computer graphics, and RECCE's learning approach based on reconstruction networks and identity features may better generalize to these two forgery methods. Figure 5 presents the average AUC results of training on one forgery method and testing on the remaining three forgery methods. Overall, the method proposed in this chapter demonstrates good generalization ability in dealing with unseen forgery types.

Table 4. Cross-manipulation evaluation in terms of AUC (%) on FaceForensics++ (HQ) with four manipulation methods.

Train Set	Methods	Test Set				AVG
		DF	FS	F2F	NT	
DF	Freq-SCL [21]	98.91	66.87	58.90	63.61	63.13
	MAT [29]	99.51	67.33	66.41	66.01	66.58
	RECCE [25]	99.65	74.29 *	70.66	67.34	70.76
	OUR	99.74 *	65.84	72.15 *	76.57 *	78.58 *
FS	Freq-SCL [21]	75.90	98.37	54.64	49.72	60.09
	MAT [29]	82.33	98.82	61.65	54.79	66.26
	RECCE [25]	82.39 *	98.82	64.44	56.70 *	67.84
	OUR	76.74	99.56 *	71.59 *	55.11	75.75 *
F2F	Freq-SCL [21]	67.55	55.35	93.06	66.66	63.19
	MAT [29]	73.04	65.10	97.96	71.88	70.01
	RECCE [25]	75.99	64.53	98.06	72.32 *	70.95
	OUR	78.38 *	77.57 *	98.88 *	62.07	79.22 *
NT	Freq-SCL [21]	79.09	53.99	74.21	88.54	69.10
	MAT [29]	74.56	60.90	80.61	93.34	72.02
	RECCE [25]	78.83	63.70 *	80.89 *	93.63	74.47
	OUR	88.45 *	59.43	74.39	98.20 *	80.12 *

The highest results are highlighted with *.

Figure 5. Comparative experimental results of cross-manipulation evaluations.

4.2.3. Cross-Dataset Evaluations

To further ascertain the generalization performance of ReLAF-Net across unseen data, the model was trained exclusively on the FF++ (HQ) dataset and then subjected to cross-dataset evaluations on Celeb-DF, DFDC, and DFD datasets, utilizing AUC (%) as the evaluation metric. The summarized results are presented in Table 5. While the Xception baseline model exhibited suboptimal generalization capabilities, our proposed ReLAF-Net showed commendable performance across all datasets not encountered during the training phase, even with Xception as the backbone. Despite being marginally outperformed by MLR and IID on Celeb-DF, and by IID on DFDC, ReLAF-Net maintained solid results. Notably, ReLAF-Net attained the highest performance on the DFD dataset, emphatically confirming the robust generalization proficiency of our proposed model when confronted with unfamiliar data. This remarkable performance is predominantly due to the strategic incorporation of restricted self-attention for local relationship learning and the meticulous mining of fine-grained frequency features.

Table 5. Cross-dataset generalization results of frame-level AUC (%) by training on FaceForensics++ (HQ) dataset.

Methods	FaceForensics++	Celeb-DF	DFDC	DFD
Xception [3]	96.30	36.19	48.98	87.86
SPSL [17]	96.91	76.88	66.16	-
MAT [29]	99.80	67.44	-	-
SIA [23]	96.94	77.35	-	-
IID [27]	99.32	83.00	81.23 *	93.92
BOF [15]	99.36	78.26	-	-
LRL [11]	99.46	78.26	76.53	89.24
MRL [12]	96.18	83.58 *	71.53	-
SFDG [48]	99.53	75.83	73.64	88.00
OUR	99.72 *	82.16	80.35	94.33 *

The highest results are highlighted with *.

In addition, it should be noted that when the model trained on FF++ is tested on different datasets, the detection efficacy varies. Specifically, the generalization performance is high for DFD, but it progressively declines for Celeb-DF and DFDC. This variation is primarily attributed to inherent differences in data quality, the deepfake generation techniques utilized in each dataset, and dataset-specific biases. Unlike FF++, the datasets DFD, Celeb-DF, and DFDC employ advanced forgery methods, the details of which remain undisclosed. Moreover, Celeb-DF and DFDC often implement post-processing techniques to diminish the signs of forgery. Additionally, DFDC represents more realistic scenarios with faces presented at various angles and under different lighting conditions. Consequently, when conducting tests across datasets, the AUC score decreases in response to these factors.

4.2.4. Robustness to Perturbations

In addition to excellent generalization over unknown data, it is essential for detectors to withstand the common damages that videos may experience on social media. We examined the effect of degrees of H.264 video compression on detector performance by training the detector with FF++ c23 (i.e., HQ version) and subsequently testing it at video compression rates c23 and c40. Figure 6 illustrates that although Face X-ray exhibits strong generalization capabilities, it is highly sensitive to compression, suggesting that hybrid boundaries are vulnerable. The F^3-Net delivered results that were slightly below ours but demonstrated significant enhancement over two comparative methods, indicating that incorporating frequency information effectively alleviates compression effects. Moreover, we evaluated our method against the Xception backbone by measuring the AUC at six distinct levels of three specific image perturbations: color saturation changes, Gaussian blur, and block-wise distortion, following the settings outlined in reference [49]. As depicted in Figure 7,

our method showed superior robustness to the backbone network, particularly under Gaussian blur scenarios where Xception nearly failed as interference intensified, whereas our method remained effective. This advantage stems from our method's reliance on detecting block-level correlations rather than visual features. However, this also resulted in weaker robustness of our method under block-wise distortion compared to the backbone. It is crucial to note that in practical scenarios, such block damage is significantly less frequent than the other two perturbations, thus maintaining our method's considerable benefits.

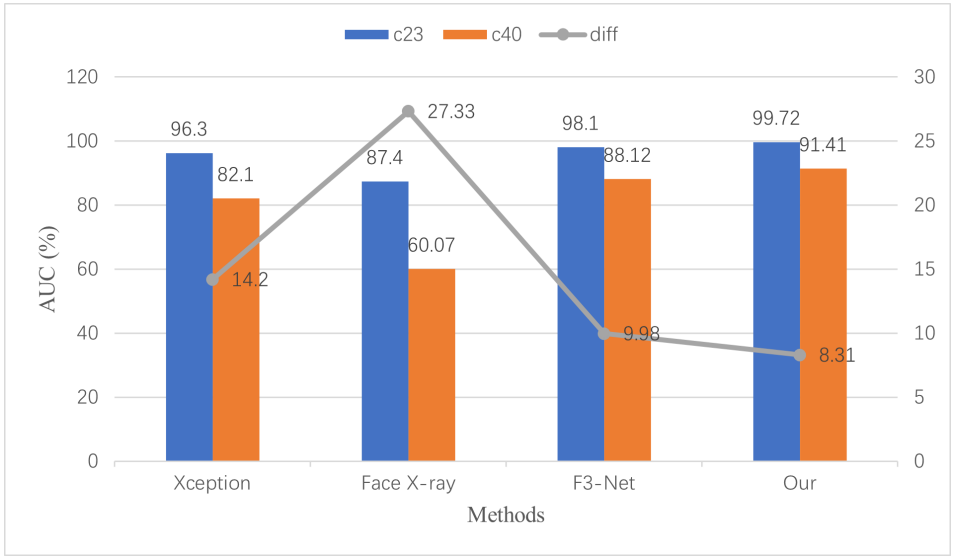


Figure 6. Robustness to compression.

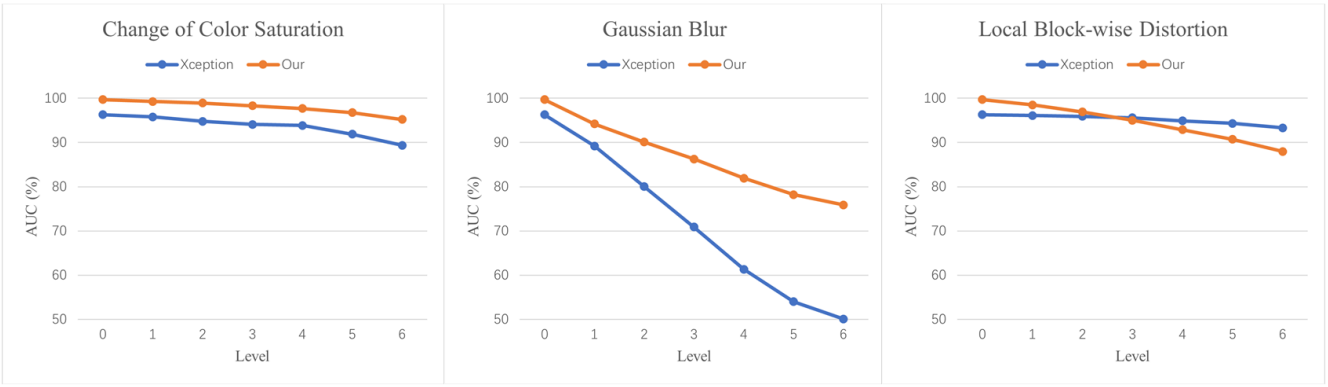


Figure 7. Robustness to common perturbations.

4.3. Ablation Study

For the three components proposed, we designed six different combinations to conduct ablation studies, to evaluate the contribution of each component to the model's performance.

- No.1: BASE only (the check mark indicates that this component is included), which refers to using only the baseline model Xception for testing on the RGB branch;
- No.2: Adds ReLAM on the basis of BASE;
- No.3: Uses the baseline model with ALFEM for detection in the frequency domain;
- No.4: Adds ReLAM on the basis of ALFEM;
- No.5: Uses a dual-branch network paired with both DCFFM and ALFEM;
- No.6: Incorporates the full model with all three proposed components.

As shown in Table 6, by comparing the results of Experiment No.1 and Experiment No.2, it can be seen that when only the baseline model BASE is used, the accuracy and AUC of the model are 95.25% and 96.93%, respectively. However, when the ReLAM module

is introduced on top of BASE, the accuracy and AUC of the model increase to 96.61% and 99.19%, respectively, proving the significant improvement of the ReLAM module on the overall performance of the model. Similarly, by comparing the results of Experiment No.3 and Experiment No.4, the effectiveness of the ReLAM module is also confirmed on the frequency domain branch. Furthermore, by comparing the results of Experiments No.1, No.3, No.5, and Experiments No.2, No.4, No.6, it is found that under the same configuration, the performance of the dual-branch structure model surpasses that of the single-branch model. Finally, when all modules (BASE, ALFEM, DCFFM, and ReLAM) are combined, the model's accuracy and AUC reach the highest at 97.92% and 99.72%, respectively, fully proving that the combination of these four modules can achieve the best model performance. The comparison results of the ROC curves for different component models are shown in Figure 8.

Table 6. Abalation study on the influence of different components on FaceForensics++ (HQ) dataset.

Number	BASE	ALFEM	DCFFM	ReLAM	ACC (%)	AUC (%)
1	✔				95.25	96.93
2	✔			✔	96.61	99.19
3		✔			95.32	97.44
4		✔		✔	96.73	99.27
5	✔	✔	✔		96.39	99.02
6	✔	✔	✔	✔	97.92	99.72

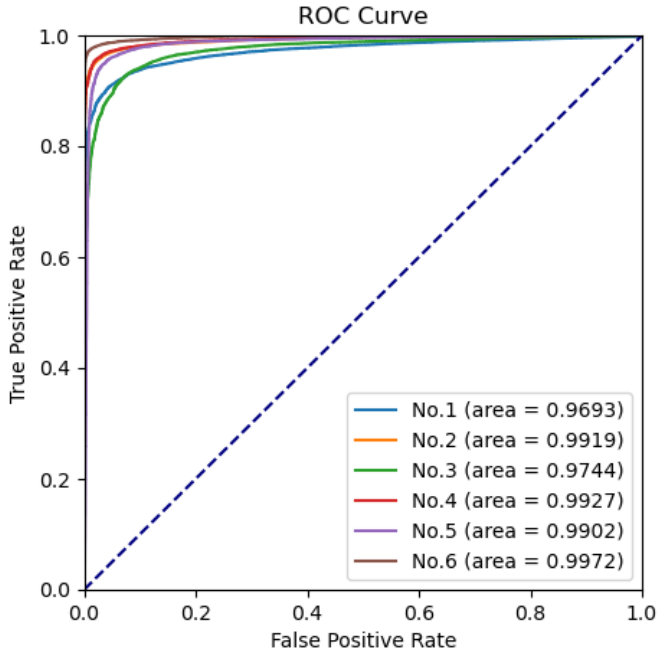


Figure 8. Ablation experiment results plotted using ROC curve.

4.4. Results Display

To demonstrate the actual detection efficacy of our model on input face images, we selected several face sequences from the test sets of FF++, DFD, Celeb-DF, and DFDC, and recorded the prediction scores from the final binary classification feed-forward layer of the model. In this system, a real face is labeled as 0, while a fake face is labeled as 1. As illustrated in Figure 9, the classifier's output distinctly differentiates between real and fake faces in the FF++ and DFD datasets. Although the score discrepancy is reduced for the Celeb-DF and DFDC samples, the classifier still effectively decides the authenticity of the video.

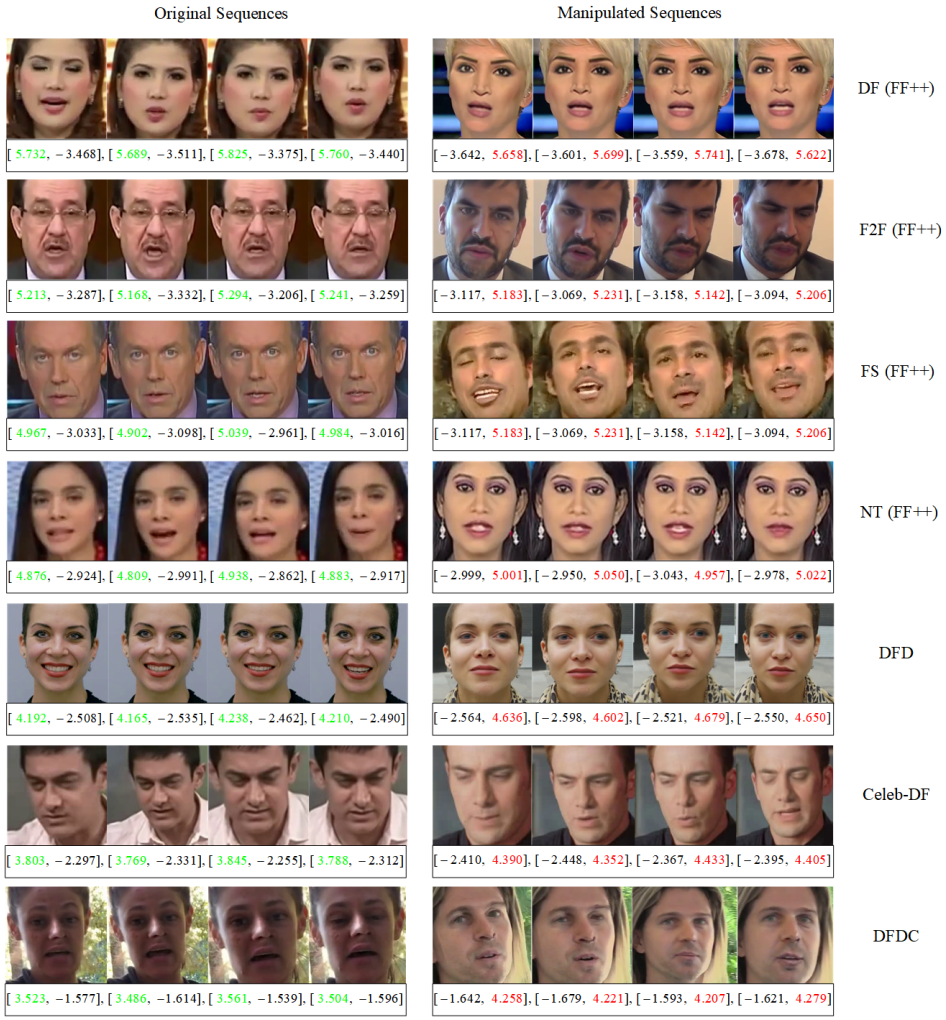


Figure 9. Binary classification prediction scores of our model on input sequence samples of FF++, DFD, Celeb-DF and DFDC datasets.

5. Conclusions

In summary, this study introduces a novel model that refines localized attention features through multi-scale relationships in the spatial-frequency domain, significantly enhancing deepfake detection. By integrating a restricted self-attention mechanism and a fine-grained frequency feature extraction algorithm, the model substantially improves generalizability and the ability to detect subtle forgery traces. This approach effectively addresses the challenge of identifying sophisticated deepfake manipulations and demonstrates superior performance in accurately recognizing altered images and videos compared to existing methods. However, it is crucial to acknowledge the limitations of our model. Although effective in various testing scenarios, its performance may fluctuate with extremely high-quality deepfakes or those generated by novel methods. Moreover, the computational complexity associated with refining localized attention features may restrict its applicability in real-time detection systems.

Despite these limitations, our method offers significant potential for use in areas that require robust verification of digital media authenticity, such as security surveillance, media forensics, and content moderation on social platforms. Additionally, the three improvement modules we propose can be integrated with various model architectures in the field of computer vision to address similar challenges in their respective tasks.

Further research and adaptation of this model could play a crucial role in protecting digital content integrity and combating the spread of misinformation through deepfakes. Future studies should explore the feature patterns of various forgery methods, optimize

the model's computational efficiency, and enhance its robustness through the application of data augmentation and adversarial training strategies.

Author Contributions: Conceptualization, Y.G. and P.Z.; methodology, Y.Z.; software, Y.Z.; validation, Y.G., Y.Z. and P.Z.; formal analysis, Y.M.; data curation, Y.Z.; writing—original draft preparation, Y.Z.; writing—review and editing, P.Z.; supervision, Y.G.; funding acquisition, Y.G. and Y.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the China Postdoctoral Science Foundation OF FUNDER grant number 2019M650608, the National Social Science Foundation of China OF FUNDER grant number 19ZDA127, and the National Natural Science Foundation of China OF FUNDER grant number 61772047.

Informed Consent Statement: Not applicable.

Data Availability Statement: This research employed publicly available datasets for its experimental studies. The FaceForensics++ and DFD datasets can be obtained by visiting https://github.com/ondyari/FaceForensics (accessed on 28 September 2022), and the Celeb-DF dataset can be obtained by visiting https://github.com/yuezunli/celeb-deepfakeforensics (accessed on 1 November 2022), and the DFDC dataset can be obtained by visiting https://www.kaggle.com/c/deepfake-detection-challenge/data (accessed on 1 March 2023).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GAN	Generative Adversarial Network
ReLAFNet	Refining Localized Attention Features Network
ALFEM	Adaptive Local Frequency Extraction Module
DCFFM	Dual-Channel Feature Fusion Module
ReLAM	Restricted Local Attention Module
DF	DeepFakes
FS	FaceSwap
F2F	Face2Face
NT	NeuralTextures

References

1. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *NIPS* **2014**, *27*, 2672-2680.
2. Afchar, D.; Nozick, V.; Yamagishi, J.; Echizen, I. Mesonet: A compact facial video forgery detection network. In Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security, Hong Kong, China, 11–13 December 2018; pp. 1–7.
3. Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October 2019–2 November 2019; pp. 1–11.
4. Masi, I.; Killekar, A.; Mascarenhas, R.M.; Gurudatt, S.P.; AbdAlmageed, W. Two-branch recurrent network for isolating deepfakes in videos. In Proceedings of the Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 667–684.
5. Chai, L.; Bau, D.; Lim, S.N.; Isola, P. What makes fake images detectable? understanding properties that generalize. In Proceedings of the Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 103–120.
6. Dong, S.; Wang, J.; Ji, R.; Liang, J.; Fan, H.; Ge, Z. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 3994–4004.
7. Deepfakes Github. Available online: https://github.com/deepfakes/faceswap (accessed on 20 March 2024).
8. Faceswap Github. Available online: https://github.com/MarekKowalski/FaceSwap (accessed on 20 March 2024).
9. Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C.; Nießner, M. Face2face: Real-time face capture and reenactment of rgb videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2387–2395.
10. Thies, J.; Zollhöfer, M.; Nießner, M. Deferred neural rendering: Image synthesis using neural textures. *ACM Trans. Graph.* **2019**, *38*, 1–12. [CrossRef]

11. Chen, S.; Yao, T.; Chen, Y.; Ding, S.; Li, J.; Ji, R. Local relation learning for face forgery detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 2–9 February 2021; Volume 35, pp. 1081–1088.
12. Yang, Z.; Liang, J.; Xu, Y.; Zhang, X.Y.; He, R. Masked relation learning for deepfake detection. *IEEE Trans. Inf. Forensics Secur.* **2023**, *18*, 1696–1708. [CrossRef]
13. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
14. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
15. Miao, C.; Chu, Q.; Li, W.; Gong, T.; Zhuang, W.; Yu, N. Towards generalizable and robust face manipulation detection via bag-of-feature. In Proceedings of the 2021 International Conference on Visual Communications and Image Processing, Munich, Germany, 5–8 December 2021; pp. 1–5.
16. Frank, J.; Eisenhofer, T.; Schönherr, L.; Fischer, A.; Kolossa, D.; Holz, T. Leveraging frequency analysis for deep fake image recognition. In Proceedings of the 37th International Conference on Machine Learning, Virtual, 13–18 July 2020; pp. 3247–3258.
17. Liu, H.; Li, X.; Zhou, W.; Chen, Y.; He, Y.; Xue, H.; Yu, N. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 772–781.
18. Li, L.; Bao, J.; Zhang, T.; Yang, H.; Chen, D.; Wen, F.; Guo, B. Face X-ray for more general face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5001–5010.
19. Zhao, T.; Xu, X.; Xu, M.; Ding, H.; Xiong, Y.; Xia, W. Learning self-consistency for deepfake detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 15023–15033.
20. Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; Shao, J. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In Proceedings of the Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Volume 12357, pp. 86–103.
21. Li, J.; Xie, H.; Li, J.; Wang, Z.; Zhang, Y. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6458–6467.
22. Gu, Q.; Chen, S.; Yao, T.; Chen, Y.; Ding, S.; Yi, R. Exploiting fine-grained face forgery clues via progressive enhancement learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 22 February–1 March 2022; Volume 36, pp. 735–743.
23. Sun, K.; Liu, H.; Yao, T.; Sun, X.; Chen, S.; Ding, S.; Ji, R. An information theoretic approach for attention-driven face forgery detection. In Proceedings of the Computer Vision-ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; Volume 13674, pp. 111–127.
24. Cozzolino, D.; Rössler, A.; Thies, J.; Nießner, M.; Verdoliva, L. Id-reveal: Identity-aware deepfake video detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 15108–15117.
25. Cao, J.; Ma, C.; Yao, T.; Chen, S.; Ding, S.; Yang, X. End-to-end reconstruction-classification learning for face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4113–4122.
26. Dong, X.; Bao, J.; Chen, D.; Zhang, T.; Zhang, W.; Yu, N.; Guo, B. Protecting celebrities from deepfake with identity consistency transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 9468–9478.
27. Huang, B.; Wang, Z.; Yang, J.; Ai, J.; Zou, Q.; Wang, Q.; Ye, D. Implicit identity driven deepfake face swapping detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 4490–4499.
28. Dang, H.; Liu, F.; Stehouwer, J.; Liu, X.; Jain, A.K. On the detection of digital face manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5781–5790.
29. Zhao, H.; Zhou, W.; Chen, D.; Wei, T.; Zhang, W.; Yu, N. Multi-attentional deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2185–2194.
30. Fei, J.; Dai, Y.; Yu, P.; Shen, T.; Xia, Z.; Weng, J. Learning second order local anomaly for general face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 20270–20280.
31. Shang, Z.; Xie, H.; Zha, Z.; Yu, L.; Li, Y.; Zhang, Y. PRRNet: Pixel-Region relation network for face forgery detection. *Pattern Recognit.* **2021**, *116*, 107950. [CrossRef]
32. Rao, Y.; Ni, J.; Xie, H. Multi-semantic CRF-based attention model for image forgery detection and localization. *Signal Process.* **2021**, *183*, 108051. [CrossRef]
33. Wang, G.; Jiang, Q.; Jin, X.; Li, W.; Cui, X. MC-LCR: Multimodal contrastive classification by locally correlated representations for effective face forgery detection. *Knowl.-Based Syst.* **2022**, *250*, 109114. [CrossRef]
34. Wang, C.; Deng, W. Representative forgery mining for fake face detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14923–14932.
35. Zhu, X.; Wang, H.; Fei, H.; Lei, Z.; Li, S.Z. Face forgery detection by 3d decomposition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2928–2938.

36. Wang, R.; Juefei-Xu, F.; Ma, L.; Xie, X.; Huang, Y.; Wang, J.; Liu, Y. Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces. *arXiv* **2019**, arXiv:1909.06122.
37. Luo, Y.; Zhang, Y.; Yan, J.; Liu, W. Generalizing face forgery detection with high-frequency features. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 16317–16326.
38. Sun, Z.; Han, Y.; Hua, Z.; Ruan, N.; Jia, W. Improving the efficiency and robustness of deepfakes detection through precise geometric features. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3609–3618.
39. Zheng, Y.; Bao, J.; Chen, D.; Zeng, M.; Wen, F. Exploring temporal coherence for more general video face forgery detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 15044–15054.
40. Hu, J.; Liao, X.; Liang, J.; Zhou, W.; Qin, Z. Finfer: Frame inference-based deepfake detection for high-visual-quality videos. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 22 February–1 March 2022; Volume 36, pp. 951–959.
41. Jung, T.; Kim, S.; Kim, K. Deepvision: Deepfakes detection using human eye blinking pattern. *IEEE Access* **2020**, *8*, 83144–83154. [CrossRef]
42. Qi, H.; Guo, Q.; Juefei-Xu, F.; Xie, X.; Ma, L.; Feng, W.; Zhao, J. Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 4318–4327.
43. Haliassos, A.; Vougioukas, K.; Petridis, S.; Pantic, M. Lips don't lie: A generalisable and robust approach to face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 5039–5049.
44. Li, Y.; Yang, X.; Sun, P.; Qi, H.; Lyu, S. Celeb-df: A large-scale challenging dataset for deepfake forensics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3207–3216.
45. Dufour, N.; Gully, A. Contributing data to deepfake detection research. *Google AI Blog.* **2019**. Available online: https://research.google/blog/contributing-data-to-deepfake-detection-research (accessed on 20 March 2024).
46. Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; Ferrer, C.C. The deepfake detection challenge (dfdc) dataset. *arXiv* **2020**, arXiv:2006.07397.
47. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
48. Wang, Y.; Yu, K.; Chen, C.; Hu, X.; Peng, S. Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7278–7287.
49. Jiang, L.; Li, R.; Wu, W.; Qian, C.; Loy, C.C. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 2889–2898.