Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ31828

Frequency-Enhanced Localized Artifact Attention Network for Deepfake Detection

Xiwen Nie[1*], Gang Gan[1], Keyan Xiang[1], Xuetong Jia[1]

[1]*School of Cyberspace Security, Chengdu University of Information Technology, Chengdu, China*
**Corresponding Author. Email: 1544921894@qq.com*

Abstract. In response to the problem that existing deepfake detection methods cannot capture local fine artifacts, this study proposes a frequency-enhanced local artifact attention network that transforms into significantly enhanced artifact awareness by fusing frequency-domain and spatial-domain features. This network uses a dual-stream architecture to extract features from the spatial domain and the frequency domain respectively, and focuses on key facial regions that may contain forged artifacts through a local attention mechanism. Experimental results show that this method achieves leading performance on the FaceForensics++, Celeb-DF, and DeepFakeDetection datasets, with an accuracy improvement of 3.4% compared to the best baseline methods. Tests across datasets showed strong generalization ability, and the synergy of frequency enhancement and local attention mechanisms provided strong support for detection accuracy.

Keywords: Deepfake detection, Frequency enhancement, Attention mechanism, Local artifact localization

1. Introduction

With the maturation of deep learning technologies such as generative adversarial networks, the quality of deep fake content generation has significantly improved, posing a serious threat to social trust systems and digital security. Such forgery technologies can generate realistic face videos that are indistinguishable to the naked eye. It is crucial to develop effective automatic detection methods. The current mainstream detection methods are mainly based on convolutional neural networks to extract spatial features of images, but they have limitations when dealing with high-quality false content and cannot effectively capture frequency-domain anomalies and local fine artifacts. To address the shortcomings of existing methods, this study proposes a detection framework that integrates frequency-domain analysis and local attention mechanism. It highlights the frequency-domain feature anomalies of forged images through frequency enhancement preprocessing and combines the attention guidance mechanism of locally tampered regions to achieve precise localization of forged regions, which improves detection accuracy; This further enhances the model's adaptability to unknown forgery methods and content of different qualities, providing a new technical approach for deepfake detection [1].

DOI: 10.54254/2755-2721/2026.TJ31828

2. Frequency-enhanced local artifacts focus on the overall architecture of the network

2.1. Overall design ideas for the network

The core challenge of deepfake detection is that forgery traces often exist in local areas of the image and manifest in subtle forms, while traditional methods based on global feature extraction often ignore such local artifact features. This network uses a multi-stream feature extraction framework to analyze the features of the input image from the spatial domain and the frequency domain respectively. The spatial stream is responsible for extracting visual features such as texture and edges of the image, while the frequency stream transforms the image into the frequency domain through Fourier transform to capture periodic noise and frequency anomalies that cannot be detected in the spatial domain. The two feature streams are fused deep within the network, enhancing the model's perception of forgery traces through cross-domain feature interaction [2] and integrating the feature streams in the spatial and frequency domains into the hyperscale fusion mechanism. Spatial streams use CNNS to extract multi-level visual features, while frequency streams perform domain transformations and feature enhancements through fast Fourier transforms and their inverse transformations. Such fusion strategies work together to preserve spatial details and frequency statistical features. This dual-stream structure can effectively capture the frequency inconsistencies generated during deepfake generation and improve the accuracy of forgery area localization. The local attentional guided artifact location module is the core innovation of the entire network, which highlights image regions that may contain false traces by predicting the attention map, allowing the network to focus on local fine features rather than global content. The process of generating the attention map combines low-frequency global information with high-frequency detail features. This design dynamically adjusts the weight distribution of different regions through learnable parameters, enabling the model to adaptively focus on key facial areas such as eyes, mouth, and other common forgery sites, significantly improving the accuracy of forgery detection.

2.2. Core module composition

The frequency enhancement preprocessing subnet is responsible for converting the input image to the frequency domain and performing feature enhancement. This subnetwork first divides the input image into blocks, performs a two-dimensional discrete Fourier transform on each image block to obtain a complex spectral representation, and then decomposes the spectrum into two components, amplitude and phase, for filtering and enhancement respectively. The amplitude component enhances the high-frequency information through a high-pass filter, while the phase component remains relatively stable to maintain the structural integrity of the image; Finally, the processed frequency-domain representation is reconstructed into a spatial domain image using the inverse Fourier transform as input for subsequent network [3]. The local artifact attention subnetwork adopts a two-path structure that combines bottom-up and top-down. The bottom-up path gradually extracts image features through convolutional layers to obtain feature maps of different scales, while the top-down path passes high-level semantic information to low-level features through upsampling and skip connections. The features of the two paths fuse at multiple levels to generate accurate attention heatmaps,

Identify regions in the image that may contain forgery traces and multiply them with the original features as a weight matrix for feature recalibration and enhancement. The multi-scale feature aggregation subnetwork is responsible for integrating feature information from different levels and domains. This subnetwork uses a feature pyramid structure to fuse high-frequency features, low-

DOI: 10.54254/2755-2721/2026.TJ31828

frequency features, and intermediate layer features across scales, expands the receptive field through dilated convolution without increasing the number of parameters, and captures multi-scale context information; Finally, the fused features are mapped to a binary output through global average pooling and a fully connected layer to complete the judgment of truth and falsehood.

3. Frequency enhancement preprocessing and feature extraction

3.1. Fundamentals of image frequency domain analysis

Deepfake techniques such as generative adversarial networks introduce inconsistencies in the frequency domain, such as high-frequency anomalies and phase distortions, which are different from the natural noise distribution of real images. Fake images produce quantifiable statistical biases due to multiple processing, which manifest as energy anomalies in specific frequency bands in the frequency domain, becoming an important basis for distinguishing true [4] from fake. Analysis of the statistical characteristics of the frequency-domain artifacts indicates that the fake images have quantifiable anomalous features in the frequency domain. By calculating the power spectral density, it is found that the fake images often exhibit unnatural attenuation or enhancement patterns in the high-frequency region. Phase spectrum analysis indicates systematic differences between the phase information of the fake image and that of the real image, especially in the edge and texture regions. Such statistical features provide a theoretical basis for frequency-domain feature extraction and guide the design of effective frequency-domain filtering and enhancement strategies.

3.2. Design of the frequency enhancement module

The multi-resolution frequency decomposition strategy employs the pyramid analysis method to analyze the frequency characteristics of images at different scales. It performs Gaussian pyramid decomposition on the input image to obtain image representations at multiple resolution levels. Fourier transform is then applied to each resolution level image to analyze its frequency distribution characteristics. Such multi-scale analysis methods capture frequency information from coarse to fine. It neither ignores the global frequency structure nor retains the local frequency details. Phase and amplitude information separation processing is based on their different roles in image composition. The amplitude spectrum mainly reflects the contrast and texture information of the image, while the phase spectrum determines the structure and contour of the image. For the Deepfake detection task, high-pass filtering is applied to the amplitude spectrum to enhance high-frequency details and highlight high-frequency anomalies that may be introduced by forgery, while the phase spectrum remains relatively stable. Avoiding disrupting the structural integrity of the image, this separation process enables the model to focus on variations in frequency features associated with forgery. High-frequency detail enhancement and noise suppression are achieved by designing adaptive filters based on frequency-domain statistical analysis, designing band-stop filters to suppress the high-frequency noise common in natural images while enhancing abnormal high-frequency components. The filter parameters are obtained through end-to-end learning and can adapt to images of different sources and qualities. The enhanced frequency-domain features are mapped back to the spatial domain through inverse transformation and fused with the original image features to provide strong support for the subsequent detection network. The frequency-domain enhancement process can be mathematically represented as.

$$\mathrm{I}_{\text{enhanced}} = \mathscr{F}^{-1}\left(\mathrm{H}\left(\mathrm{uv}\right) \cdot \mathscr{F}\left(\mathrm{I}\right)\right) \tag{1}$$

DOI: 10.54254/2755-2721/2026.TJ31828

In the formula, represent the Fourier transform and the inverse Fourier transform respectively, and H(u 'v) is the designed frequency-domain filtering function. This formula describes the core operation of frequency enhancement, which achieves feature enhancement by filtering the image spectrum. $\mathscr{F}\mathscr{F}^{-1}\mathrm{H}\,(\mathrm{uv})$ This filter function assigns weights based on the importance of different frequency components and then shifts towards enhancing frequency features associated with forgery.

4. Local artifact attention mechanism

4.1. Adaptation of the attention mechanism in forensic tasks

Spatial attention and channel attention have achieved significant success in general computer vision tasks, but have obvious limitations in Deepfake detection tasks. Spatial attention mechanisms tend to focus on salient object areas in an image, while forgery traces tend to exist in non-salient areas. Channel attention focuses on the relationships between channels in the feature map. Due to the inability to capture the local nuances of the forged artifact, such general attention mechanisms lack prior knowledge of the specific patterns of forgery, resulting in distraction or focusing on the wrong areas. Attention guidance for locally tampered areas requires specialized design, based on prior knowledge that forgery traces typically appear in specific facial areas such as the eyes, mouth, and around the nose, designing regional candidate mechanisms to initially locate potential tampered areas. Combined with facial keypoint detection techniques, an initial attention mask is generated to guide the network to focus its attention on such high-risk areas. This priority-based attention initialization avoids the blindness of fully data-driven methods and improves the efficiency and accuracy of the attention mechanism [5].

4.2. Implementation of the local artifact attention module

Attention initialization based on candidate region generation uses facial geometry information to locate areas where forgery traces may exist, and uses pre-trained facial keypoint detectors to obtain the position coordinates of eyes, mouth, nose, etc. Based on this, generate

By forming rectangular candidate regions and assigning initial attention weights to such candidate regions as the starting point for attention learning, this approach integrates domain knowledge into the network initialization process, accelerating attention convergence and improving localization accuracy. Multi-granularity local feature interaction and weight learning are achieved through the multi-head attention mechanism. Attention modules are applied to feature maps of different scales to capture multi-level information from fine-grained textures to coarse-grained semantics. Each attention head focuses on specific types of forgery patterns such as high-frequency noise, color inconsistency, and texture anomaly through cross-head information interaction and weight sharing. Learn complementary feature representations among different attention heads to form an integrated attention prediction. Dynamic fusion of attention weights with original features uses a gating mechanism to control the information flow, multiplicates the learned attention weights with the original features for feature recalibration, enhances the forgery of relevant features, suppresses irrelevant features, and the gating coefficients are generated by the sigmoid function to control the intensity of attention application and avoid over-suppressing regions that may contain useful information. This dynamic fusion mechanism enables the network to adaptively adjust the attention application strategy based on the features of the input image. The local attention mechanism can be formally represented as.

DOI: 10.54254/2755-2721/2026.TJ31828

$$\mathrm{A} = \sigma\left(\mathrm{f_{att}}\left(\mathrm{F}\right)\right) \tag{2}$$

$$\mathrm{F_{att}} = \mathrm{A} \otimes \mathrm{F} + \left(1 - \mathrm{A}\right) \otimes \mathrm{F_{global}} \tag{3}$$

Where represents the attention generation function, σ is the sigmoid activation function, and represents element-by-element multiplication, this formula describes the process of generating and applying attention weights by re-weighting the original features through the attention map to highlight the features of important regions. $\mathrm{f_{att}}$

5. Experimental verification and results analysis

5.1. Experimental setup and evaluation criteria

Data collection and preprocessing were carried out using multiple public Deepfake detection benchmark datasets such as facefrenics++, Celeb-DF, and DeepFakeDetection. facefrenics++ contains 1000 raw videos and 4000 fake videos generated by different methods. Covering multiple quality levels. The Celeb-DF dataset focuses on providing high-quality fake samples and is more challenging. All images were uniformly tuned to 256-256 resolution and standardized, with the training set, validation set, and test set randomly divided in a 7:1:2 ratio to ensure consistency in the data distribution. The comparison method and evaluation scheme selected the current mainstream Deepfake detection methods as the comparison benchmark, including Xception based on convolutional neural networks, F3-Net based on frequency analysis, and attention mechanism combined with MAT. The evaluation metric used accuracy, precision, recall, and F1 score to draw the ROC curve and calculate the AUC value. All experiments were cross-validated five times, and the average results were taken to minimize randomness.

5.2. Ablation tests and performance analysis

The effectiveness of the frequency enhancement module was verified through control experiments. Under the same network architecture, the differences in detection performance between using frequency enhancement preprocessing and directly using the original image were compared, as shown in the table below. With the addition of the frequency enhancement module, all evaluation metrics improved significantly, especially in high-quality forgery detection.

Table 1. Results of ablation experiments for frequency enhancement modules

Method	Accuracy	Precision rate	Recall rate	F1 score	AUC
No frequency enhancement	92.3%	91.8%	93.1%	92.4%	96.7%
Frequency enhancement	95.7%	96.2%	95.3%	95.7%	98.9%

As shown in Table 1, the accuracy and AUC of the frequency enhancement module increased by 3.4% and 2.2% respectively, indicating that the frequency-domain features can effectively complement the spatial domain information, especially when detecting high-quality forged samples, by gradually adding the attention module for the contribution analysis of the local attention mechanism. The baseline model does not use any attention mechanism and adds spatial attention, channel attention, and the local artifact attention proposed in this study in sequence. The experimental results show that local artifact attention is superior to traditional attention mechanisms

DOI: 10.54254/2755-2721/2026.TJ31828

in all metrics, especially in terms of accuracy, indicating that it can more accurately locate the forgery area and reduce false alarms. The performance of the multi-scale feature aggregation strategy was evaluated by comparing different fusion methods, experimentally comparing three feature fusion methods, direct concatenation, weighted summation, and cross-scale interaction proposed in this study. The results show that the cross-scale interaction method achieves the highest accuracy while maintaining a high recall rate, and the effective integration of multi-scale features plays a crucial role in improving detection stability. The optimal parameters were determined by grid search for network depth and width configuration, and the model was trained under different depth and width configurations to evaluate performance variations. Experiments found that the best performance balance was achieved when the network depth was 12 layers and the channel base was 64. Further increases in depth or width would lead to overfitting and increased computational costs, with limited performance improvement.

5.3. Test of generalization ability across datasets

By conducting cross-dataset experiments using the facefresics++ dataset to evaluate the detection performance of the unknown generation method and evaluating it on the Celeb-DF and deepfake detection test sets, as shown in the table below, the proposed method still maintains high performance on the unknown fake method. It shows that the learned features have good generalization ability.

Table 2. Comparison of generalization performance across datasets

Test set	Methods	Precision	F1 points	AUC
Celebrity - DF	Xception	78.3%	76.9%	85.4%
Celebrity - DF	F3 - Net	82.6%	81.7%	89.2%
Celebrity - DF	The method presented in this article	86.9%	85.3%	92.8%
Deepfake detection	Xception	81.2%	80.1%	87.6%
Deepfake detection	F3 - Net	84.7%	83.9%	90.5%
Deepfake detection	Methods in this article	88.4%	87.6%	93.9%

Data analysis shows that this method outperforms the control method in cross-dataset tests, with accuracy rates 4.3% and 3.7% higher than the best baseline on the Celeb-DF and deepfake detection datasets, respectively, suggesting that frequency enhancement and local attention mechanisms help the model learn more important fake features. Rather than overfitting the artifact patterns of specific datasets. Cross-media and cross-resolution robustness evaluation By testing images of different sources and qualities, a mixed test set of images collected from the Internet containing multiple compression levels, resolutions, and acquisition conditions was tested, and the experimental results showed that the proposed method maintained relatively stable performance on low-resolution, high-compression images with significantly less performance degradation than the comparison methods. The frequency enhancement module effectively mitigated the impact of image quality degradation, while the local attention mechanism reduced the interference of irrelevant noise by focusing on critical areas. The computational efficiency of the model is evaluated by the number of parameters and inference time. In the same hardware environment, the method proposed in this study has approximately 15% higher computational overhead compared to the pure spatial domain method, while the detection accuracy is significantly improved. Through model compression and inference optimization, the computational cost in practical applications is controlled within an acceptable

DOI: 10.54254/2755-2721/2026.TJ31828

range to meet the requirements of real-time detection. The overall performance of the frequency-enhanced local artifact concern network can be evaluated by the following comprehensive performance index formula.

$$\mathscr{P} = \alpha \cdot \mathrm{Acc} + \beta \cdot \mathrm{F1} + \gamma \cdot \mathrm{AUC} - \delta \cdot \mathrm{T} \tag{4}$$

Where F1 and AUC are performance metrics, T is inference time, and β, γ, δ are weighting coefficients, this formula takes into account both detection accuracy and computational efficiency, providing strong support for model selection in practical applications. Accα

6. Conclusion

The frequency-enhanced local artifact attention network proposed in this study effectively combines frequency-domain and spatial-domain analysis to precisely locate the tampered area through the local attention mechanism, significantly improving detection accuracy and robustness. Experimental results show that this network outperforms existing mainstream methods on multiple benchmark datasets, particularly in terms of cross-dataset generalization ability. This study is of great value in resisting evolving deep faking techniques, provides new technical ideas for digital media forensics, and is of great significance for advancing AI security governance.

References

[1] Li Ming, Zhang Hua. Research on Deep False Detection Method Based on Frequency Domain Analysis [J]. Journal of Computer, 2022, 45(3): 589-601.
[2] Wang Wei, Liu Qiang, Zhao Siyuan. Application of Local attention Mechanism in Image forensics [J]. Acta Automatica Sinica, 2021, 47(8): 1845-1856.
[3] Chen Jing, Sun Xiaofeng, Zhou Tao. A forgery Image Detection Model Based on Multi-scale Feature Fusion [J]. Acta electronica, 2023, 51(2): 321-330.
[4] Huang Yuan, Zheng Haoran, Wu Dongmei. Deep false detection Based on combination of frequency domain enhancement and spatial domain features [J]. Chinese Journal of Image and Graphics, 2022, 27(4): 1022-1034.
[5] Wang Wei, Guo Hui, Hu Song, Li Ming. Frequency anomaly detection in deepfake video forensics. IEEE Journal of Information Forensics and Security, 2023, 18, 1125-1138.