BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 21/25 pp889–904
DOI: 10.26599/BDMA.2024.9020030
Volume 7, Number 3, September 2024

Local Region Frequency Guided Dynamic Inconsistency Network for Deepfake Video Detection

Pengfei Yue, Beijing Chen*, and Zhangjie Fu

Abstract: In recent years, with the rapid development of deepfake technology, a large number of deepfake videos have emerged on the Internet, which poses a huge threat to national politics, social stability, and personal privacy. Although many existing deepfake detection methods exhibit excellent performance for known manipulations, their detection capabilities are not strong when faced with unknown manipulations. Therefore, in order to obtain better generalization ability, this paper analyzes global and local inter-frame dynamic inconsistencies from the perspective of spatial and frequency domains, and proposes a Local region Frequency Guided Dynamic Inconsistency Network (LFGDIN). The network includes two parts: Global SpatioTemporal Network (GSTN) and Local Region Frequency Guided Module (LRFGM). The GSTN is responsible for capturing the dynamic information of the entire face, while the LRFGM focuses on extracting the frequency dynamic information of the eyes and mouth. The LRFGM guides the GTSN to concentrate on dynamic inconsistency in some significant local regions through local region alignment, so as to improve the model's detection performance. Experiments on the three public datasets (FF++, DFDC, and Celeb-DF) show that compared with many recent advanced methods, the proposed method achieves better detection results when detecting deepfake videos of unknown manipulation types.

Key words: deepfake video detection; dynamic inconsistency; local region; local region frequency

1 Introduction

In recent year, with the continuous development and advancement of deep learning technology and deepfake generation techniques, various face-swapping applications have emerged, such as Deepfakes[1], DeepfaceLab[2], FaceSwap[3], etc. While enriching people's cultural and entertainment life, these applications have also been abused by malicious individuals, causing many social problems and posing great challenges to politics, justice, criminal investigation, reputation protection, and even social stability. In 2020, CCTV News reported that criminals used AI face-swapping to cheat face recognition systems for account and financial fraud. During the Russia-Ukraine conflict in 2022, a generated video depicting Ukrainian President Zelenskyy calling on Ukrainian soldiers to lay down their weapons was widely spread on the Internet, causing significant turmoil in the situation. Negative news about deepfake facial images and videos continues to emerge, leading countries to pay more attention to the management and control of deepfakes. In September 2018, the European Union issued the "Code of Conduct against

• Pengfei Yue, Beijing Chen, and Zhangjie Fu are with Engineering Research Center of Digital Forensics affiliated with Ministry of Education, and also with School of Computer Science, and also with Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET), Nanjing University of Information Science and Technology, Nanjing 210044, China. E-mail: yuepengfe1@163.com; nbutimage@126.com; fzj@nuist.edu.cn.
* To whom correspondence should be addressed.
Manuscript received: 2024-01-10; revised: 2024-04-05; accepted: 2024-05-07

Disinformation" to combat online rumors. In 2020, China implemented the "Management of Internet Audio and Video Information Services" regulation, whose Article 11 explicitly prohibits the production, dissemination, and distribution of false news using new technologies based on deep learning and virtual reality. In November 2023, at the first AI Security Summit hosted by the United Kingdom, 28 countries and the European Union signed the "Bletchley Declaration", containing discussions on fake news control.

In addition to the strict legislative control of deepfake in various countries, more and more research is devoted to controlling the malicious abuse of deepfake through deepfake video detection. These studies can be divided into two categories according to whether they utilize inter-frame information or not: intra-frame methods[4–9] and inter-frame methods[10–18]. Intra-frame methods first classify video frames (usually key frames) using the detection methods for deepfake image, and then make comprehensive decisions on the detection results of these frames to obtain the final video detection results. These methods detect deepfake by exploring facial manipulations and artifacts present in facial frames. Consequently, the features learned by most intra-frame methods are highly correlated with manipulation methods in the training set, resulting in easy overfitting of known manipulations and poor generalization to unknown manipulations. To overcome this problem, inter-frame methods are proposed. Since existing deepfake video technologies usually do not add restrictions on the temporal dimension, inter-frame methods exploit the inconsistencies among multiple frames for deepfake video detection. These methods reveal the inter-frame inconsistencies unrelated to manipulation methods through in-depth exploration from different perspectives, such as temporal inconsistencies[10], optical flow changes[11], and dynamic characteristics[13], thus showing stronger generalization ability than the intra-frame methods. However, existing inter-frame methods still have some shortcomings: (1) They either do not focus on some significant local regions (such as eyes, mouth, etc.), or they do not consider them comprehensively enough; (2) When extracting dynamic inconsistent features, they mainly consider performing it in the spatial domain, without the consideration of important frequency domain.

To solve the above problems, this paper proposes a Local region Frequency Guided Dynamic Inconsistency Network (LFGDIN) for deepfake video detection. In addition to global face, this model also takes into account the significant local regions of the eyes and mouth. It uses the frequency features of these local regions to guide global face network to learn dynamic inconsistencies for deepfake video detection, allowing it to obtain more accurate and general detection results.

The main contributions of this paper are as follows:

- Proposing the LFGDIN for deepfake video detection. The LFGDIN uses the frequency features of highly discriminative local regions to guide the global face network to effectively learn the dynamic inconsistencies among frames.
- Designing a Local Region Frequency Guidance Module (LRFGM) for guiding the global face network. It includes two stages: frequency feature extraction and local region guidance. The former effectively extracts the frequency dynamic information of significant local regions. The latter generates an interest attention map to guide the global spatiotemporal features to pay more attention to the significant local regions.
- Compared with multiple benchmark models, the proposed network is verified its superiority, especially in terms of generalization performance.

2 Related Knowledge

2.1 Deepfake manipulation and deepfake detection

(1) Deepfake technology

Deepfake is a type of image and video manipulation technique based on deep learning, which can be divided into four types of face manipulation: entire face synthesis, attribute manipulation, identity swap, and facial reenactment. Entire face synthesis usually uses Generative Adversarial Network (GAN)[19] to create completely non-existent faces. Attribute manipulation also uses GAN to modify certain attributes of the face, such as skin or hair color, age, and gender, etc. Identity Swap replaces a source face in an image or video with a target face, commonly using techniques such as Deepfakes[1] and FaceSwap[3]. Facial reenactment modifies the facial expressions of source face in an image or video, commonly using techniques, such as Face2Face[20] and NeuralTextures[21]. As shown in Fig. 1, the typical deepfake image manipulation process is divided into three stages: (1) face extraction; (2) face manipulation;

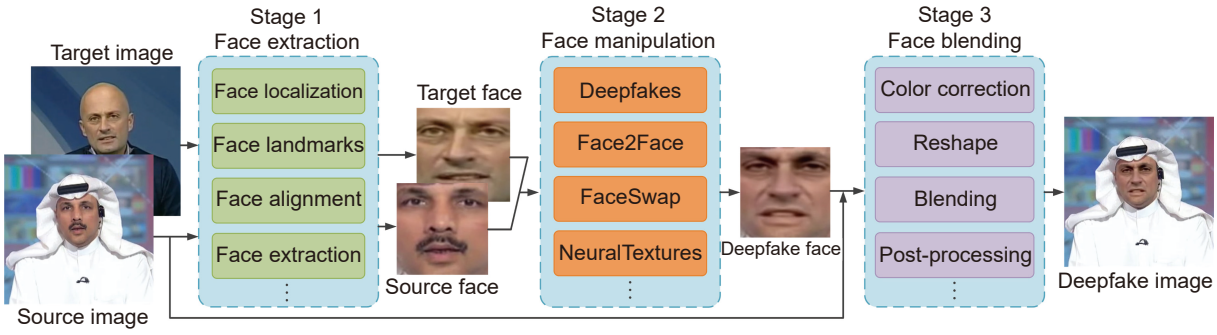


Fig. 1 Workflow of deepfake image manipulation.

and (3) face blending.

In the first stage, the process of face extraction can be expressed as

$$x_\tau = E\left(D\left(i_\tau\right)\right) = E\left(c_\tau\right), \tau \in \{t, s\} \tag{1}$$

where i_τ represents input image; x_τ represents face; c_τ represents face coordinates. When τ is t, i_t, x_t, and c_t represent input target image, target face, and target face coordinates, respectively; when τ is s, i_s, x_s, and c_s represent input source image, source face and source face coordinates, respectively. $D(\,)$ is face localization operation, and $E(\,)$ is face extraction operation. The face localization and preprocessing operations are performed on i_t and i_s to obtain target face coordinates c_t, and source face coordinates c_s. Then, the face extraction operations are applied based on c_t and c_s to obtain target face x_t and source face x_s.

In the second stage, the process of face manipulation can be expressed as

$$x_f = M\left(x_t, x_s\right) \tag{2}$$

where $M(\,)$ is a face manipulation method, such as Deepfakes, FaceSwap, Face2Face, etc. This process applies the required face manipulation method to the target face x_t and source face x_s to obtain manipulated face x_f.

In the third stage, the process of face blending can be expressed as

$$i_f = B\left(i_s, c_s, x_f\right) \tag{3}$$

where $B(\,)$ is a face blending operation. This process performs deformation operations and color correction on the manipulated face x_f, and blends x_f into the source image i_s based on the source face coordinates c_s. Then, some post-processing operations are performed to obtain final deepfake image i_f.

The manipulation process of deepfake videos is similar to that of deepfake images. The difference is that it first splits the video into different frames, then performs deepfake image manipulation on each frame, and finally combines all the manipulated deepfake frames to obtain the deepfake video.

(2) Deepfake detection technology

Deepfake detection refers to determine whether the given image or video is deepfake or not through analyzing manipulation traces. For deepfake image detection, since deepfake images mostly blend target faces into source images and there are some differences between the target faces and source images, some manipulation traces appear. As shown in Fig. 2a, there are some obvious manipulation traces in the deepfake facial images, such as visible boundaries, shape distortion, facial blur, and color differences. These

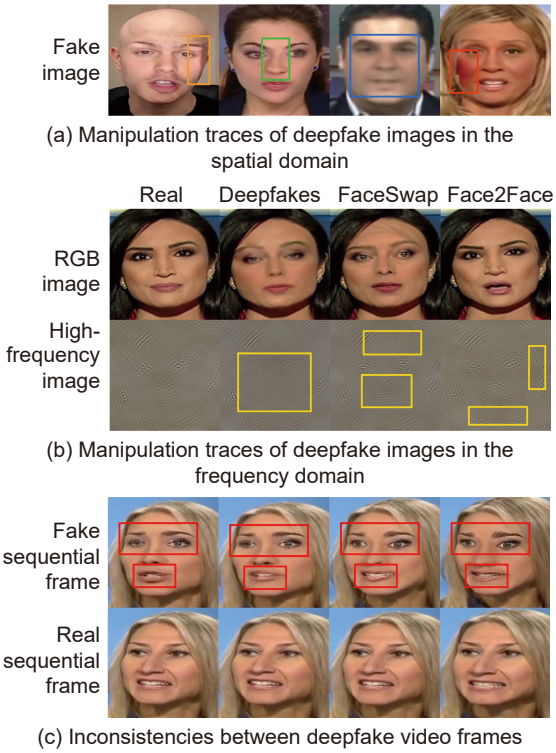


Fig. 2 Traces of manipulation in fake images and videos.

traces are often caused by post-processing operations in the deepfake manipulation process. As shown in Fig. 2b, deepfake facial images not only have obvious artifacts in the spatial domain, but also have traces of manipulation in the frequency domain. Based on these manipulation traces, many Convolutional Neural Networks (CNNs) for deepfake image detection have been proposed. The detection process is shown in Fig. 3. It can be roughly divided into two stages: face extraction and deepfake face detection.

In the first stage, the process of face extraction can be expressed as

$$x = E\left(D\left(i\right)\right) \tag{4}$$

where i represents the input image. This stage performs face localization on the input image, and then extracts the facial image x.

In the second stage, the process of deepfake face detection can be expressed as

$$p = \mathrm{IF}\left(x\right),\ p \in \{0,\ 1\} \tag{5}$$

where IF ( ) represents deepfake facial image detector. The facial image x is fed into the detector IF for detection to obtain the final prediction result. If p is 1, it indicates that the input is a deepfake image, otherwise it is a real image.

Since most of the deepfake videos are generated by merging deepfake images generated frame by frame, they often exhibit inter-frame inconsistencies, such as different facial expression changes and facial organ movements, as shown in Fig. 2c. Based on these inconsistencies, the deepfake video detection process can also be roughly divided into two stages: face extraction and deepfake face detection.

In the first stage, the required number of frames are extracted from the video to be detected. The process of this stage can be represented as

$$x_\mu = E\left(D\left(j_\mu\right)\right),\ \mu \in \{1, 2, \ldots,\ n\} \tag{6}$$

where j represents input video frame. In this stage, the face localization is performed on n input frames, of the video, and then n facial images x_μ are extracted.

In the second stage, the process of deepfake face detection can be expressed as

$$p = \mathrm{VF}\left(x_1,\ , x_2, \ldots,\ x_n\right),\ p \in \{0,\ 1\} \tag{7}$$

where VF ( ) represents deepfake facial video detector. n facial images x_μ are fed into the detector VF to obtain the final prediction result. If p is 1, it indicates that the input is a deepfake video, otherwise it is a real video.

2.2 Related work on deepfake video detection

In order to mitigate the security threats that deepfake videos may pose, researchers have proposed many methods for deepfake video detection. These methods can be divided into two categories based on whether they utilize inter-frame information or not: intra-frame methods and inter-frame methods.

Intra-frame methods typically decompose each video into frames to explore the authenticity discrimination features within a single frame, then classify each frame (usually the key frames) as real or fake, and finally integrate the results of all the considered frames to obtain the final video detection result. Most of these methods attempt to explore subtle manipulation traces in deepfake facial frames from the spatial domain or frequency domain. Lin et al.[4] proposed a deepfake detection method that combines multi-scale features and Vision Transformer, which can effectively capture face details and global information at different scales. Zhao et al.[5] proposed a method based on a multi-attention mechanism, which effectively captures the texture information of local regions and has good intra-dataset detection performance for high-quality frame images. Li et al.[6] proposed the method Face X-ray in consideration of the blending boundaries existing in forged frames. Li et al.[7] proposed an adaptive

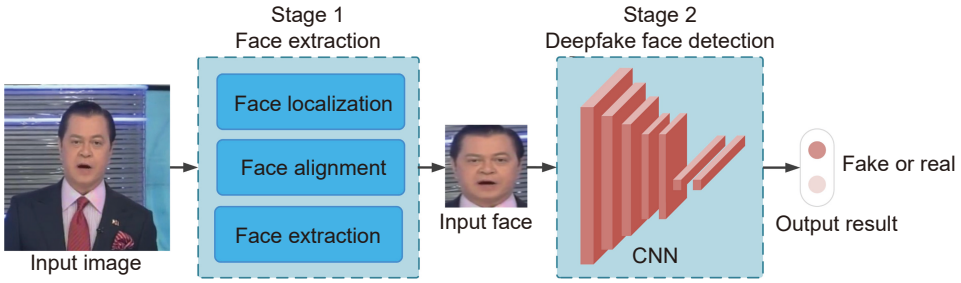


Fig. 3 Workflow of deepfake image detection.

frequency feature generation module to extract frequency features, and introduced a single center loss function to reduce the intra-class difference and enlarge the inter-class difference. This method has excellent intra-dataset performance on various datasets with different compression qualities. Qian et al.[8] proposed F3-Net composed of two frequency-aware branches. This method also has excellent intra-dataset performance in challenging low-quality frame image detection. However, these intra-frame methods have good performance in intra-dataset detection tasks but have limited generalization ability to unknown manipulations. This is because they learn discriminative features that are highly related to the manipulation methods in the training set, resulting in overfitting to known manipulations.

Inter-frame detection methods explore forgery features by analyzing inconsistencies between multiple frames and use them to detect deepfake videos. Since video manipulations often result in inconsistencies between frames, and these features are often independent of the manipulation methods, the inter-frame methods can better address the problem of overfitting to known manipulations of the intra-frame methods and improve the generalization ability to detect unknown manipulated videos. Liu et al.[10] utilized the temporal inconsistencies between video frames for deepfake detection. They trained a model using GRU and triplet loss, and their method demonstrates good generalization ability on unknown datasets. Caldelli et al.[11] proposed using optical flow to capture temporal inconsistencies along the video timeline, and their method also exhibits good generalization ability. Since deepfake algorithms are typically trained on face images with open eyes, Saealal et al.[12] used blink frequency for deepfake facial video detection. Wang et al.[13] utilized the facial dynamic inconsistencies between frames in deepfake videos, fusing local dynamic information from the lips and global dynamic information from the entire face through complementary cross-dynamic fusion. Ding et al.[18] explored short-term inter-frame inconsistencies using multiple RGB video frames and interactively fused reconstructed frames based on the frequency domain phase, thereby better capturing spatiotemporal inconsistencies for deepfake video detection. However, existing methods that utilize inter-frame inconsistencies for detection still have some limitations. Firstly, these methods either do not focus on significant local regions (such as eyes and mouth) or do not consider them comprehensively. The experiments in Ref. [22] demonstrate that eyes and mouth are the most indicative regions for the detection of authenticity in facial local regions. Therefore, it is crucial to consider these significant local regions comprehensively for deepfake detection. Secondly, current inter-frame methods primarily focus on spatial domain when extracting dynamic inconsistent features, without incorporating important frequency domain information. Frequency features can effectively reveal manipulation artifacts in frame images from a frequency domain perspective.

3 Method

The structure of the LFGDIN proposed in this paper is shown in Fig. 4. It mainly consists of two parts: GSTN and LRFGM. The two parts handle the global region (entire face) and the local regions (eyes and mouth), respectively. The LRFGM extracts local frequency features to guide GSTN in paying more attention to the dynamic changes in significant local regions.

3.1 GSTN

The goal of GSTN is to capture the dynamic information and the long-range dependency relationship of the global face in the video. In GSTN, UniformerV2[23] based on Vision Transformer (ViT)[24] is employed as the backbone, combined with the Region Of Interest (ROI) attention maps obtained from LRFGM to guide the model. This guidance allows GSTN to pay more attention to the dynamic inconsistency of significant local regions.

Regarding the backbone of GSTN, UniformerV2 inherits the advantages of ViT and effectively captures the global dependencies within frames. Additionally, with the help of its global UniBlock, UniformerV2 can learn long-range dependency relationship across frames, thereby enhancing the network's understanding and representation capability of videos. Specifically, as shown in Fig. 4, the global UniBlock consists of Dynamic Position (DynPos) encoding, Multi-Head Relation (MHRel) aggregator, and Feed-Forward (FeedFwd) network. Firstly, the DynPos encoding uses 3D convolution to encode positional information, allowing for better learning of global spatiotemporal features by maintaining spatiotemporal order. Secondly, the MHRel aggregator employs a learnable

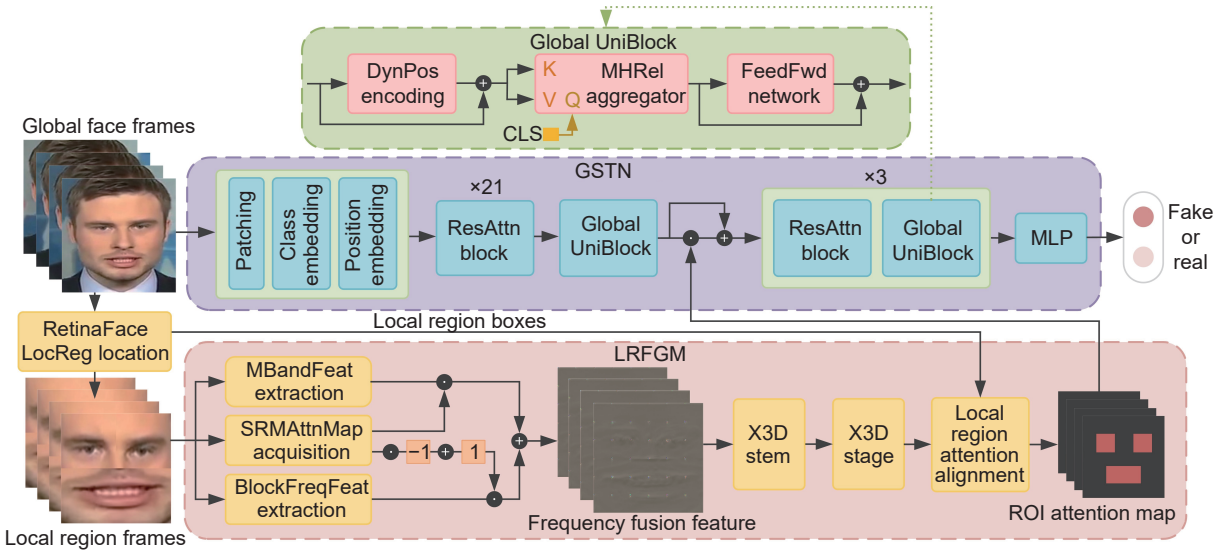


Fig. 4 Network architecture of the proposed LFGDIN.

classification token (namely CLS) as the query and uses multi-head cross-attention to model long-range dependencies between the query and global spatiotemporal features, converting the query into a video representation. Finally, the FeedFwd network including two linear layers enhances the video representation.

3.2 LRFGM

Due to the importance of capturing dynamic inconsistency in local regions for deepfake video detection, LRFGM is designed to explore the inter-frame dynamic inconsistencies of local regions from the perspective of frequency domain. Certainly, within the entire network, the GSTN focusing on global facial dynamics plays a primary role, and LRFGM is an auxiliary, guiding GSTN to focus more on dynamic information in local regions. The structure of LRFGM is illustrated in Fig. 4. It takes T local region frames, each of which composes of three spliced local regions, as input. These local regions are obtained by localizing and cropping the left eye, right eye, and mouth in the facial frames using the RetinaFace[25] network. Subsequently, the local region frames go through two stages: frequency feature extraction and local region guidance. In the former stage, the frequency fusion features of each local region frame are first extracted; then, the frequency dynamic inconsistencies of local regions between frames are explored using 3D convolutions. In the latter stage, the global spatiotemporal features are guided using the local region frequency dynamic features with the help of local region bounding boxes.

3.2.1 Frequency feature extraction stage

As shown in Fig. 4, for each local region frame $e_u \in \mathbf{R}^{3\times H\times W}$, $u \in \{1, 2, \ldots, T\}$, where H and W represent the height and width of the local region frame, respectively the stage starts with Frequency Fusion Feature Extraction (FFFE), followed by learning frequency dynamic information using 3D convolutions to obtain the local region frequency dynamic features. The FFFE uses Steganalysis Rich Model (SRM)[26] Attention Map (SRMAttnMap) to fuse Multi-Band Features (MBandFeat) and Blockwise Frequency Features (BlockFreqFeat). Taking the first local region frame $e_1 \in \mathbf{R}^{3\times H\times W}$ as an example, the specific process of relevant features extraction and attention map acquisition is shown in Fig. 5.

The purpose of MBandFeat extraction is to comprehensively and adaptively mine manipulation traces from different frequency bands to obtain rich frequency perception clues. As shown in Fig. 5a, it first employs three binary filters $h_k \in \mathbf{R}^{3\times H\times W}$, $k \in \{1, 2, 3\}$, to divide the Discrete Cosine Transform (DCT) spectrum into three frequency bands: low, medium, and high. h_1 extracts the first 1/16 of the spectrum to obtain the low-frequency band information; h_2 extracts the portion between 1/16 and 1/8 of the spectrum to obtain the medium-frequency band information; h_3 extracts the remaining of the spectrum to obtain the high-frequency band information. The spectra of the three frequency bands can be obtained by

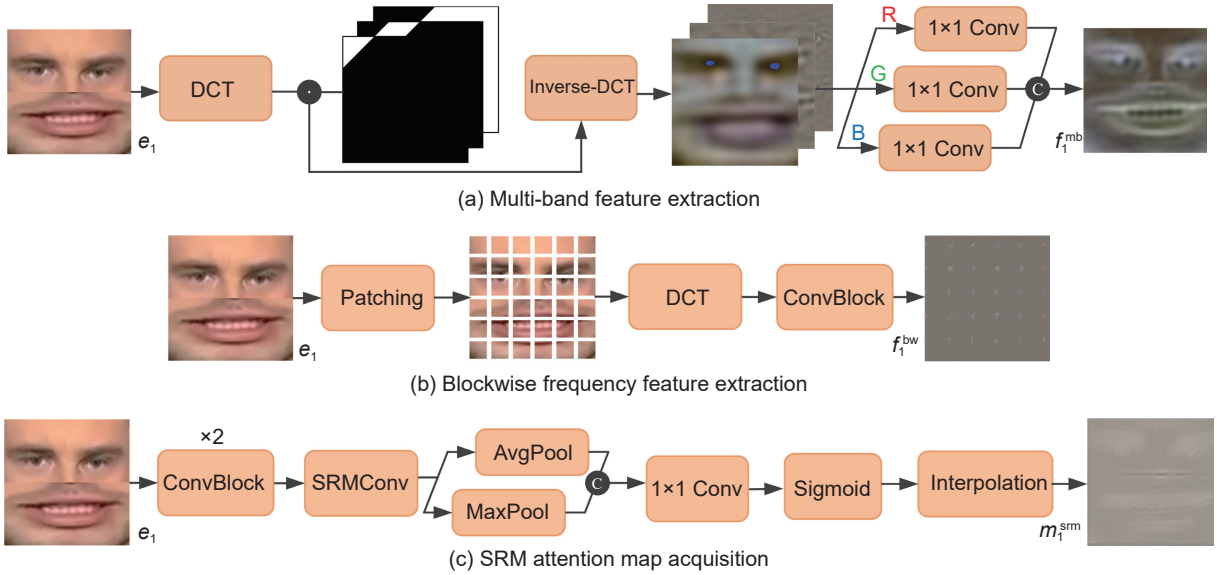


Fig. 5 Frequency feature extraction and SRM attention map acquisition of the first local region frame e_1.

$$s_k = \mathrm{DCT}(e_1) \odot (h_k + l_k),\ k \in \{1, 2, 3\} \tag{8}$$

where "$\odot$" is element-wise multiplication and $l_k \in \mathbf{R}^{1\times H\times W}$ is the learnable weight normalized to $[-1, 1]$. By adding learnable weights to each of the three filters, these filters can adaptively extract information from different frequency bands in the spectrum. Then, the spectra of three different frequency bands are fused to obtain multi-band features $f_1^{\mathrm{mb}} \in \mathbf{R}^{3\times H\times W}$ for the first local region frame. The specific operations are as follows:

$$\begin{aligned} f_1^{\mathrm{mb}} &= V(\mathrm{DCT}^{-1}(s_1), \mathrm{DCT}^{-1}(s_2), \mathrm{DCT}^{-1}(s_3)) = \\ & V(z_1, z_2, z_3) = \\ & [\mathrm{convR}(z_1, z_2, z_3), \mathrm{convG}(z_1, z_2, z_3), \mathrm{convB}(z_1, z_2, z_3)] \end{aligned} \tag{9}$$

where $\mathrm{DCT}^{-1}(\,)$ represents the inverse DCT operation; $z_k \in \mathbf{R}^{3\times H\times W}$, $k \in \{1,2,3\}$, represent the images of the low, medium, and high-frequency bands, respectively, obtained after performing the inverse DCT operation on the spectrum; $V(\,)$ is the operation of multi-band images fusion which utilizes three 1×1 convolution operations $\mathrm{convR}(\,)$, $\mathrm{convG}(\,)$, and $\mathrm{convB}(\,)$ to fuse the R, G, and B channels of the three different frequency band images, respectively; $[\cdot]$ denotes the concatenation along the channel dimension.

The purpose of BlockFreqFeat extraction is to conduct a more detailed analysis of local regions from a spectrum perspective. As shown in Fig. 5b, it divides the local region frame into 6×6 blocks and performs DCT transform on each block; then, it concatenates the spectra of all blocks along the spatial dimension to obtain the block-wise spectrum; finally, the block-wise spectrum is fed into a convolutional block that includes convolution, layer normalization, and ReLU activation, obtaining block-wise frequency features $f_1^{\mathrm{bw}} \in \mathbf{R}^{3\times H\times W}$.

The purpose of SRMAttnMap acquisition is to highlight manipulation traces from the perspective of high-frequency noise, thereby better guiding the extraction of frequency fusion features. Research in Ref. [27] has shown that image manipulations can disrupt the consistency of the original image's noise pattern. The SRM noise, which represents high-frequency signals in the image, can not only suppress content information to extract important details but also effectively reveal inconsistencies in noise patterns unrelated to the manipulation methods, thus enhancing the model's generalization to some extent. Considering that the eye and mouth regions are relatively high-frequency regions in the face and are more prone to be manipulated, this paper uses SRM to extract high-frequency noise for guiding the extraction of frequency fusion features in these local regions. As shown in Fig. 5c, the local region frame is passed through two convolutional blocks, and three 3×3 SRM filters to extract high-frequency noise (the three SRM filters are shown in Fig. 6. Subsequently, spatial attention from Convolutional Block Attention Module (CBAM)[28] is

$$\begin{pmatrix} -1 & 2 & -1 \\ 2 & -4 & 2 \\ -1 & 2 & -1 \end{pmatrix} \begin{pmatrix} -1 & 0 & -1 \\ 0 & -4 & 0 \\ -1 & 0 & -1 \end{pmatrix} \begin{pmatrix} 2 & -1 & 2 \\ 1 & -4 & 1 \\ 2 & -1 & 2 \end{pmatrix}$$

Fig. 6 Three 3×3 SRM filters.

applied to further emphasize the manipulation traces and obtain an attention map. Finally, bilinear interpolation upsampling is used to align the attention map with the local region frame on the spatial scale, obtaining the final SRM attention map $m_1^{\mathrm{SRM}} \in \mathbf{R}^{1\times H\times W}$ of the first local region frame.

For the frequency fusion part, as shown in Fig. 4, this paper utilizes the SRM attention map to guide the acquisition of local region frequency fusion features $f_{\mathrm{fusion},\,1}^{\mathrm{lfreq}} \in \mathbf{R}^{3\times H\times W}$ for the first local region frame, which can be expressed as

$$f_{\mathrm{fusion},\,1}^{\mathrm{lfreq}} = (1 - m_1^{\mathrm{SRM}}) \odot f_1^{\mathrm{bw}} + m_1^{\mathrm{SRM}} \odot f_1^{\mathrm{mb}} \tag{10}$$

After extracting the frequency fusion feature $f_{\mathrm{fusion},\,u}^{\mathrm{lfreq}} \in \mathbf{R}^{3\times H\times W}$, $u \in \{1, 2, \ldots, T\}$ for each local region frame sequentially, the final local region frequency fusion features $f_{\mathrm{fusion}}^{\mathrm{lfreq}} \in \mathbf{R}^{3\times T\times H\times W}$ is obtained by concatenating them along the temporal dimension.

Finally, 3D convolution is used to further extract features from the local region frequency fusion features, allowing for a better exploration of the temporal inconsistencies between frames. Research in Ref. [29] has shown that (1) Vit-based models considered in our GSTN perform poorly in handling high-frequency components and local details in images because they focus on capturing global information through self-attention mechanisms; (2) CNN models can effectively capture local details and edge information corresponding to high-frequency parts in images through convolution operations. Therefore, this paper adopts the X3D network[30] based on 3D convolution for a more in-depth extraction of the local region frequency fusion feature, obtaining the local region frequency dynamic features $f_{\mathrm{dyn}}^{\mathrm{lfreq}} \in \mathbf{R}^{D\times T\times H_f\times W_f}$, where D, T, H_f, and W_f represent the number of channels, frames, height, width of the frequency dynamic features, respectively. As shown in Fig. 4, the X3D network consists of two parts: stem and stage. The X3D Stem transforms the input features into a more suitable representation for subsequent processing. The X3D Stage stacks five 3D residual blocks to progressively extract higher-level local region spatiotemporal features.

3.2.2 Local region guidance stage

Considering that the features obtained in the frequency feature extraction stage represent the frequency dynamic inconsistencies of local regions, this paper designs Local Region Attention Alignment (LRAA) operation to obtain ROI attention maps for global spatiotemporal feature guidance. The LRAA adopts the spatial attention mechanism of the CBAM to further highlight the frequency dynamics of multiple local regions, and guides the GSTN to focus on dynamic inconsistency in the significant local regions through local region alignment.

As shown in Fig. 7, in this stage, the attention maps are first sequentially extracted from the local region frequency dynamic features $f_{\mathrm{dyn}}^{\mathrm{lfreq}} \in \mathbf{R}^{D\times T\times H_f\times W_f}$ obtained in the previous stage along the temporal dimension. Taking the local region frequency feature $f_{\mathrm{dyn},\,1}^{\mathrm{lfreq}} \in \mathbf{R}^{D\times H_f\times W_f}$ of the first time sampling point on the temporal dimension as an example, the specific operation is given by

$$m_1^{\mathrm{freq}} = \mathrm{SA}\,(f_{\mathrm{dyn},\,1}^{\mathrm{lfreq}}) \tag{11}$$

where $m_1^{\mathrm{freq}} \in \mathbf{R}^{1\times H_f\times W_f}$ represents the output frequency attention map, and SA ( ) is the spatial attention operation.

Next, each local region in the frequency attention map is aligned to the global space, obtaining the final ROI attention map $m_1^{\mathrm{roi}} \in \mathbf{R}^{1\times P\times P}$, where P represents the height or width of the ROI attention map. The specific operations are as follows:

$$\begin{aligned} m_1^{\mathrm{roi}} =& \mathrm{RGSA}\,(\mathrm{RA}\,(m_1^{\mathrm{freq}}, b_1^{\mathrm{le}}),\\ & \mathrm{RA}\,(m_1^{\mathrm{freq}}, b_1^{\mathrm{re}}), \mathrm{RA}\,(m_1^{\mathrm{freq}}, b_1^{\mathrm{m}}), m^{\mathrm{z}}) =\\ & \mathrm{RGSA}\,(m_1^{\mathrm{le}}, m_1^{\mathrm{re}}, m_1^{\mathrm{m}}, m^{\mathrm{z}}) \end{aligned} \tag{12}$$

where b_1^{le}, b_1^{re}, and b_1^{m} represent the bounding boxes of the left eye, right eye, and mouth in the corresponding facial frame, respectively, and are resized through downsampling to match the spatial dimension of the global spatiotemporal features; RA ( ) denotes the ROI align operation[31], which maps the regions of the left eye, right eye, and mouth in the frequency attention map m_1^{freq} to the sizes of b_1^{le}, b_1^{re}, and b_1^{m}, respectively, obtaining the left eye attention map $m_1^{\mathrm{le}} \in \mathbf{R}^{1\times H_{\mathrm{le}}\times W_{\mathrm{le}}}$, right eye attention map $m_1^{\mathrm{re}} \in \mathbf{R}^{1\times H_{\mathrm{re}}\times W_{\mathrm{re}}}$, and mouth attention map $m_1^{\mathrm{m}} \in \mathbf{R}^{1\times H_m\times W_m}$; $m^{\mathrm{z}} \in \mathbf{R}^{1\times P\times P}$ is a zero-filled mask with the same spatial dimensions as the global spatiotemporal features; RGSA ( ) represents the ROI and global space alignment operation, which aligns the local region attention maps m_1^{le}, m_1^{re}, and m_1^{m} to the zero-filled mask m^{z}, and adds them to the m^{z}

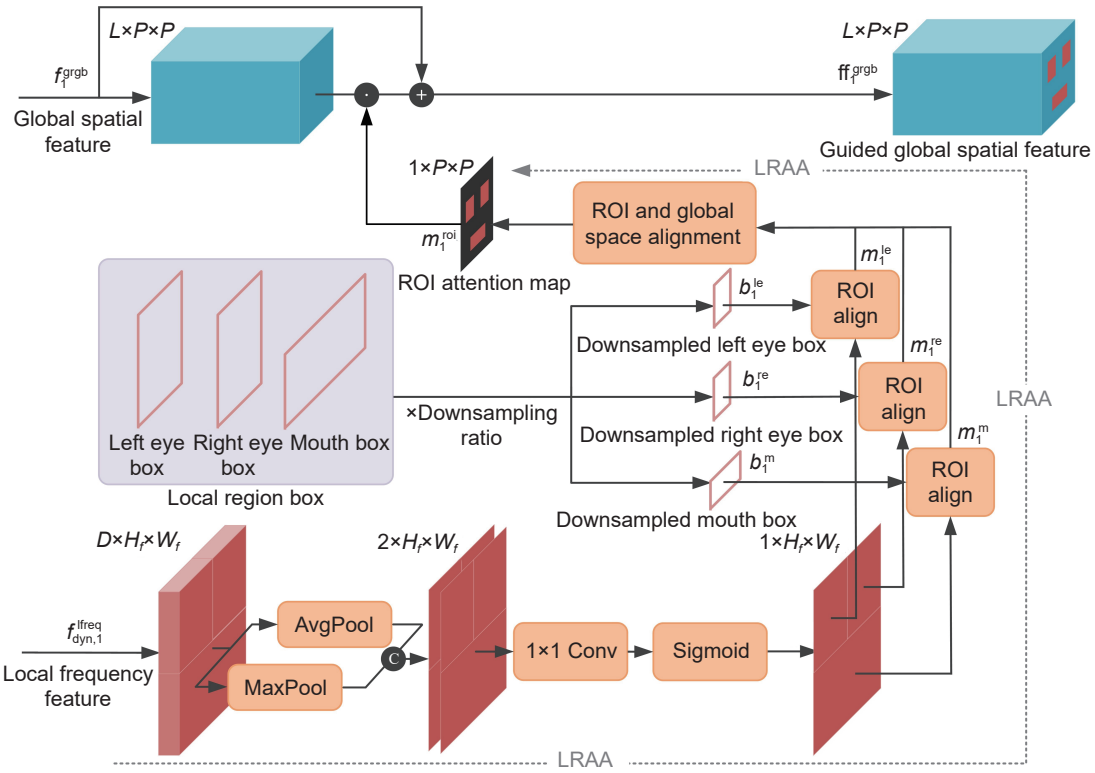


Fig. 7 Local region guidance stage of the first time sampling point on the temporal dimension.

according to the positions of the bounding boxes, obtaining the final ROI attention map m_1^{roi}.

Finally, the global spatial features of the first time sampling point on the temporal dimension are guided to pay more attention to the local regions using the ROI attention map. The specific operation is given by

$$\text{ff}_1^{\text{grgb}} = f_1^{\text{grgb}} + f_1^{\text{grgb}} \odot m_1^{\text{roi}} \tag{13}$$

where $f_1^{\text{grgb}} \in \mathbf{R}^{L\times P\times P}$ represents the global spatial features of the first time sampling point, L represents the number of channels, and $\text{ff}_1^{\text{grgb}} \in \mathbf{R}^{L\times P\times P}$ represents the global spatial features after guidance.

By applying Eqs. (11) to (13) sequentially, the global spatial features of each time sampling point along the temporal dimension are guided. These guided features are then concatenated along the temporal dimension, obtaining the guided final global spatiotemporal features $\text{ff}^{\text{grgb}} \in \mathbf{R}^{L\times T\times P\times P}$. These features reflect the dynamic inconsistent information of the global face across frames and focus more on the dynamic inconsistency of local regions.

3.3 Pseudo-code

To make the proposed method clear and easily understandable, the pseudo-code of the LFGDIN is presented in Algorithm 1. The algorithm takes global face frames, local region boxes, local region frames, and video labels as input, and the video classification result as output.

4 Experiment

4.1 Experimental setup

4.1.1 Dataset

This paper is trained on the widely used FaceForensics++ (FF++)[32] dataset and evaluated for generalization on the Celeb-DF (v2)[33], deepfake Detection Challenge (DFDC)[34], DiffFace, and DiffSwap datasets.

FF++ is a benchmark dataset that consists of 1000 real videos collected from YouTube. Four manipulation techniques, i.e, deepfakes (DF), Face2Face (F2F), FaceSwap (FS), and Neural Textures (NT), are applied to these 1000 real videos, obtaining 4000 manipulated videos. The dataset split provided by the official FF++ dataset is used, where the HQ (c23) version of FF++ is divided into 72% for training, 14% for validation, and the remaining 14% for testing. To balance the positive and negative samples in FF++, 1000 real videos are replicated three times to obtain

Algorithm 1 Procedure of training LFGDIN

Input:

T global face frames c_u, $u \in \{1, 2, \ldots, T\}$;

T local region boxes o_u; // Obtained by using RetinaFace to locate local regions within the global face frames, it is not yet downsampled

T local region frames e_u; // Obtained by cropping and splicing based on local region boxes

Label a of video

Output: Classification result y

for iter = 0 to $N_{\text{iter}} - 1$ **do**; // N_{iter} is the number of iterations

  f^{grgb} = GSTN_Stage1(c); // Extracting global spatiotemporal features from intermediate layers through GSTN

  for $u = 1$ to T **do**; // Extracting frequency fusion features for each local region frame e_u;

    Extracting multi-band feature f_u^{mb} by applying Eqs. (8) and (9) from e_u;

    Extracting block-wise frequency feature f_u^{bw} from e_u;

    Acquiring SRM attention map m_u^{SRM} from e_u;

    Extracting local region frequency fusion feature $f_{\text{fusion},u}^{\text{lfreq}}$ by applying Eq. (10);

  end for

  $f_{\text{dyn}}^{\text{lfreq}}$ = X3D ($f_{\text{fusion}}^{\text{lfreq}}$); // Extracting local region frequency dynamic features through X3D

  for $u = 1$ to T **do**; // Acquiring ROI attention map for $f_{\text{dyn},u}^{\text{lfreq}}$

    Acquiring spatial attention m_u^{freq} by applying Eq. (11);

    Acquiring ROI attention map m_u^{roi} by applying Eq. (12) based on o_u;

  end for

  Extracting guided final global spatiotemporal features ff^{grgb} by applying Eq. (13);

  y = GSTN_Stage2 (ff^{grgb}); // Acquiring classification result through GSTN

  loss = $L_{\text{bce}}(y, a)$; // Computing loss by using the binary cross-entropy loss function

  back_propagation (loss); // Computing gradients

  update (LFGDIN); // Updating the parameters of LFGDIN using AdamW

end for

4000 real videos.

All the real videos in Celeb-DF (v2) are collected from YouTube, and the celebrities appearing in these videos have different genders, ages, and races. The Celeb-DF (v2) includes 890 real videos and 5639 manipulated videos generated by improved deepfake techniques. This paper tests on 500 videos from Celeb-DF (v2), including 250 real videos and 250 fake videos.

The DFDC is a large dataset consisting of 119 197 videos, where 19 197 real videos are captured by approximately 430 actors, and the remaining 100 000 videos are generated deepfake videos from the 19 197 real videos. This paper tested 500 videos from DFDC, including 250 real videos and 250 fake videos.

To evaluate the generalization on diffusion model-based face-swapping videos, two datasets DiffFace and DiffSwap are constructed based on two diffusion models: DiffFace[35] (the first diffusion-based face-swapping model) and DiffSwap model[36] in 2023. For the DiffFace dataset, 250 real videos are selected from Celeb-DF (v2), and 250 fake videos are generated by applying the DiffFace model to the real videos through frame-by-frame manipulation. Furthermore, to enhance the quality of the fake videos, the pre-trained DiffFace model is fine-tuned on Celeb-DF (v2). The DiffSwap dataset using the DiffSwap model is also processed in the same way.

4.1.2 Evaluation metrics

This paper uses the Area Under the Curve (AUC) as the evaluation metric for the experiments. AUC refers to the area under the Receiver Operating Characteristic (ROC) curve. Assuming there are M positive samples and N negative samples, there are a total of $M \times N$ pairs of samples. Among these $M \times N$ pairs of samples, the number of cases where the positive sample

prediction probability P_p is higher than the negative sample prediction probability P_n is counted. The AUC value can be obtained using the following formula:

$$\text{AUC} = \frac{\sum I(P_p, P_n)}{M \times N},$$
$$I(P_p, P_n) = \begin{cases} 1, & P_p > P_n; \\ 0.5, & P_p = P_n; \\ 0, & P_p < P_n \end{cases} \tag{14}$$

The AUC value ranges from 0 to 1, where a higher value indicates a better classifier performance.

4.1.3 Implementation details

To consider different types of manipulation for better capturing inconsistencies between frames, and achieve a relatively stable and balanced effect, the frame selection method described in Ref. [37] is employed: 16 frames are randomly selected from each video with intervals ranging from 5 to 15 frames. In order to obtain the global face frames and local region frames, the RetinaFace[25] is used to extract faces and obtain bounding boxes for local regions from each selected frame, and the extracted faces are expanded by 30% after undergoing alignment correction and saved as the global face frame with size 224 × 224; then, the left and right eye regions are cropped from the global face frames using the bounding boxes obtained in the previous step and saved at a size of 156 × 156, while the mouth region is cropped and saved at a size of 156 × 312, after that, the three local regions of each face are spliced to form a 312 × 312 local region frame.

The PyTorch framework is used to implement the network in this paper. The UniformerV2_l14[24] and X3D_l[30] are used as the backbone networks, and their pretrained models from the Kinetics[38] dataset are used to initialize the weights. The training framework uses the AdamW optimizer with a learning rate of 1×10^{-5} and weight decay of 1×10^{-4}. The batch size is set to 4, and the model is trained for 20 epochs. During training, the network is supervised by using the binary cross-entropy loss function.

4.2 Ablation study

To evaluate the effectiveness of the two main modules introduced in this paper (FFFE and LRAA), related ablation experiments are conducted by training on FF++ (HQ) or its subset and performing intra-dataset and cross-dataset testing. The experimental results are shown in Table 1. Here four methods are as follows: (a) Only using the global GSTN without considering the local regions, i.e., without considering LRFGM that includes FFFE and LRAA; (b) Considering both the global and local regions but without considering FFFE for the local regions, i.e., using the spatial features of the local regions to guide GSTN; (c) Considering both the global and local regions but without using LRAA for the local region guidance stage, i.e., using a linear layer and activation function combination to extract frequency attention maps, without aligning each local region of the attention map to the global space, but directly adjusting the attention map size for guidance; (d) The comprehensive approach proposed in this paper, which considers both FFFE and LRAA.

As can be seen from Table 1, (1) Comparing Methods (a) and (b), it can be found that Method (b) has improved detection performance in both intra-dataset and cross-dataset which verifies that focusing on dynamic inconsistencies in local regions is effective; (2) Comparing Methods (b) and (d), it can be seen that Method (d) has further improved performance, especially in cross-dataset testing, which verifies that FFFE enables the network to analyze dynamic inconsistencies between video frames from a frequency domain perspective, improving the network's generalization ability; (3) Comparing Methods (c) and (d), Method (d) also has a better performance, verifying that using LRAA to obtain attention maps can more accurately guide GSTN to focus on dynamic information in local regions; (4) Method (d) has better detection performance than Method (a), especially in the case of cross-dataset, which proves the effectiveness of LRFGM.

Table 1 Ablation study on FF++ (HQ).

Method	FFFE	LRAA	AUC on intra-dataset					AUC on cross-dataset			
			DF	F2F	FS	NT	FF++	Celeb-DF	DFDC	DiffFace	DiffSwap
(a)	–	–	0.995	0.985	0.992	0.978	0.990	0.834	0.771	0.760	0.717
(b)	–	√	0.997	0.991	0.997	0.982	0.993	0.854	0.783	0.835	0.803
(c)	√	–	0.999	0.997	0.998	0.982	0.994	0.863	0.763	0.806	0.775
(d)	√	√	**0.999**	**0.999**	**0.999**	**0.990**	**0.997**	**0.904**	**0.808**	**0.903**	**0.857**

To better understand the decision-making processes of Methods (a), (b), (c), and (d), this article visualizes the regions they focus on for different datasets. The visualization results are shown in Fig. 8, which is obtained using Grad-CAM[39] (warm color areas represent regions that contribute significantly to the network's prediction results, while cool color areas have lower contributions). It can be seen from Fig. 8 that Method (d) pays more attention to significant local regions than Method (a), which verifies the effectiveness of the proposed LRFGM. Both Methods (b) and (d) focus on significant local regions, but Method (d) focuses more on relatively high-frequency regions in the local region due to feature extraction from the frequency domain, which verifies the effectiveness of the proposed FFFE. Comparing Methods (c) and (d), it can be seen that Method (c) makes feature weights spread across the entire face image because it does not use accurate local region alignment during local region guidance stage. In contrast, Method (d) focuses more accurately on significant local regions, which verifies the effectiveness of using the proposed LRAA for guidance. Additionally, for different forgery operations, our method can still detect manipulation traces beyond the key local regions through GSTN, such as the blended boundary in the facial contour region shown in Fig. 8 for the F2F manipulation. For NT, where manipulation is only performed on the mouth region, our method does not force the network to focus on the eye region, but rather focuses more accurately on abnormal dynamic in the mouth region.

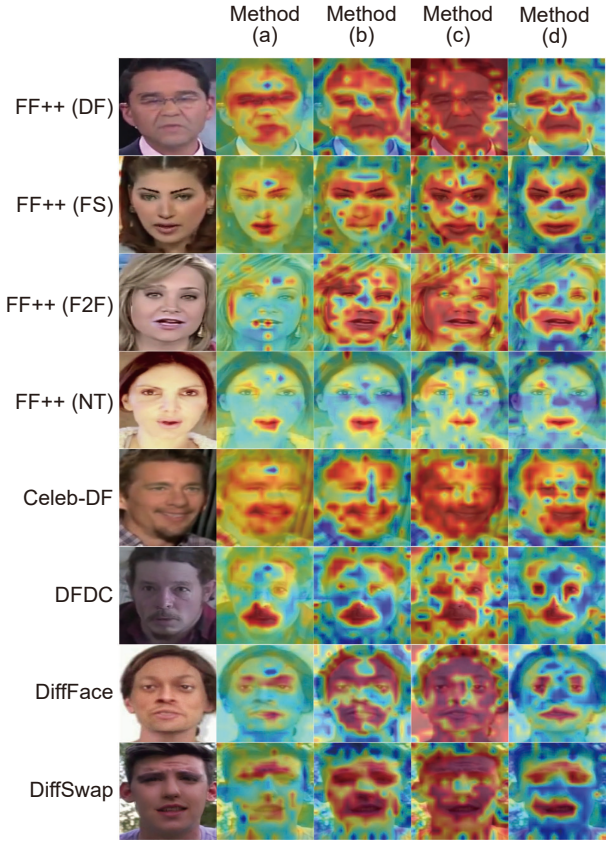


Fig. 8 Visualizing the regions of interest for the four methods on different datasets using Grad-CAM.

4.3 Comparison with recent works

To comprehensively evaluate the proposed method, nine state-of-the-art deepfake detection methods are compared, including seven intra-frame methods and two inter-frame methods. Seven intra-frame methods are as follows: (1) Xception[40] explores manipulation traces in frame images using the popular Xception network; (2) Frequency in Face Forgery Network (F3-Net)[8] mines manipulation traces in frame images through two frequency-aware branches; (3) Learning-To Weight (LTW)[41] uses meta-learning strategies to learn domain-invariant models in unknown domains; (4) Dual Contrastive Learning (DCL)[42] improves the generalization by learning contrasts between different instances and within the same instances; (5) Reconstruction-classification learning (namely RECCE)[43] enhances compact representations of real faces through reconstruction learning and classification learning; (6) Multi-Scale Frequency Contrastive Learning (MSFCL)[44] enhances the generalization through contrastive learning and multi-scale feature enhancement; (7) Domain-Invariant Feature Learning (DIFL)[45] achieves universal face forgery detection through adversarial domain generalization and center loss for learning domain-invariant features. Two inter-frame methods are as follows: (1) Spatio Temporal Inconsistency Learning and Interactive Fusion (ST-ILIF)[18] explores inconsistent information between short-term frames and combines it with reconstruction frames based on frequency domain phase; (2) Discrete Cosine Transform-based Forgery Clue Augmentation Network (FCAN-DCT)[46] detects deepfake videos by exploring spatiotemporal frequency clues among multiple frames.

4.3.1 Intra-dataset evaluation

Firstly, the intra-dataset evaluation was performed. Here the whole FF++ (HQ) dataset is used for both training and testing. As shown in Table 2, it can be

Table 2 Intra-dataset evaluation in the FF++ (HQ) dataset.

Method		AUC
Intra-frame	Xception[40]	**0.997**
	F3-Net[8]	0.986
	LTW[41]	0.991
	DCL[42]	0.993
	RECCE[43]	0.991
	MSFCL[44]	0.993
	DIFL[45]	0.993
Inter-frame	ST-ILIF[18]	0.986
	FCAN-DCT[46]	0.990
	Ours	**0.997**

found that all the compared 10 methods achieve excellent performance, no matter the intra-frame methods[8, 40–45] or the inter-frame methods[18, 46] including the proposed method. Furthermore, the proposed method obtains the best performance among 10 methods.

4.3.2 Cross-dataset evaluation

Then, the cross-dataset evaluation is carried out. In the real world, many manipulated face images are completely unknown for the manipulations. Therefore, it is crucial for the model to have cross-dataset generalization ability. The FF++ (HQ) dataset is considered for training and the other two datasets (Celeb-DF (v2) and DFDC) are used for testing.

As shown in Table 3, compared to 9 recent methods, the proposed method achieves better generalization performance in cross-dataset scenarios. The specific analysis is as follows:

(1) Regarding the Celeb-DF dataset, the proposed method demonstrates significantly better generalization performance than the other compared methods, achieving an improvement of 8.3% over FCAN-DCT, which is the best one among the 9 compared methods. The FCAN-DCT is an inter-frame method that utilizes spatiotemporal frequency inconsistencies in the video for forgery detection from a global perspective. However, it may not be sensitive enough to manipulation traces in the significant local regions. In contrast, the proposed method explores both the global dynamic inconsistencies of the entire face and dynamic inconsistency in three important local regions, obtaining strong generalization ability.

(2) Regarding the challenging DFDC dataset, the generalization abilities of all compared methods are relatively lower than that on the Celeb-DF dataset. However, the proposed method achieves a 4.7% improvement over the best method DIFL among the other 9 compared methods. Additionally, although the F3-Net employs a similar frequency feature extraction method, it exhibits lower generalization performance than the proposed method due to its intra-frame approach. In contrast, the proposed method incorporates dynamic inconsistencies between frames.

(3) For both datasets, the inter-frame methods are overall superior to the intra-frame methods in generalization performance for detecting unknown deepfake videos. This is due to the fact that most intra-frame methods primarily rely on the detection of intra-frame tampering traces that are highly correlated with manipulation techniques, while the inter-frame methods mainly focus on the detection of inter-frame inconsistencies that are unrelated to tampering techniques.

Table 3 Cross-dataset evaluation under the training on the FF++ (HQ) dataset.

Method		AUC	
		Celeb-DF	DFDC
Intra-frame	Xception[40]	0.659	0.690
	F3-Net[8]	0.732	0.701
	LTW[41]	0.771	0.746
	DCL[42]	0.823	0.767
	RECCE[43]	0.695	0.701
	MSFCL[44]	0.823	0.733
	DIFL[45]	0.799	0.772
Inter-frame	ST-ILIF[18]	0.752	–
	FCAN-DCT[46]	0.835	–
	Ours	**0.904**	**0.808**

4.3.3 Cross-manipulation evaluation

Finally, to evaluate the generalization ability to unknown manipulations, this study conducts cross-manipulation experiments on the FF++ (HQ) dataset, specifically targeting four different manipulations (DF, F2F, FS, and NT). The leave-one-out strategy is employed. Specifically, each manipulation is tested while the remaining three forgery operations are trained. The proposed method is compared to 6 recent methods because the works[8, 18, 46] do not conduct this type of experiment in their corresponding literatures. As shown in Table 4, the proposed method demonstrates the best performance on all manipulations and exhibits the best overall performance in terms of the average AUC. Specifically, the proposed method achieves an

Table 4 Cross-manipulation evaluation (AUC) in the FF++ (HQ) dataset.

Method		Manipulation				Average
		DF	FS	F2F	NT	
Intra-frame	Xception[40]	0.939	0.512	0.868	0.797	0.779
	LTW[41]	0.927	0.640	0.802	0.773	0.785
	DCL[42]	0.949	–	0.829	–	–
	RECCE[43]	0.920	0.625	0.813	0.783	0.785
	MSFCL[44]	0.941	0.656	0.814	0.792	0.801
	DIFL[45]	0.947	0.779	0.856	0.802	0.846
Inter-frame	Ours	**0.962**	**0.805**	**0.905**	**0.817**	**0.872**

improvement of nearly 3.1% average AUC over the best-performing method DIFL among 6 compared methods.

5 Conclusion

This paper proposes an LFGDIN for deepfake video detection. The LFGDIN employs ROI attention map to guide global spatiotemporal features to pay more attention to dynamic inconsistency in three significant local regions, thereby achieving strong generalization ability. The ROI attention map is obtained by performing the frequency feature extraction stage on the local region frame and conducting the local region attention alignment operation to align the local regions to global face frame. Experimental results demonstrate that the proposed method exhibits significant advantages in generalization detection tasks over several recent advanced methods. In future work, to further enhance the model's generalization performance, we plan to add the use of other inconsistent information between frames, such as depth information, noiseprint, illumination information, etc.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Nos. 62072251 and U22B2062), and the Priority Academic Program Development of Jiangsu Higher Education Institutions fund.

References

[1] M. Tora, deepfakes, https://github.com/deepfakes/faceswap/tree/v2.0.0, 2018.

[2] K. Liu, I. Perov, D. Gao, N. Chervoniy, W. Zhou, and W. Zhang, Deepfacelab: Integrated, flexible and extensible face-swapping framework, *Pattern Recognition*, vol. 141, p. 109628, 2023.

[3] M. Kowalski, FaceSwap, https://github.com/marekkowalski/faceswap, 2018.

[4] H. Lin, W. Huang, W. Luo, and W. Lu, deepfake detection with multi-scale convolution and vision transformer, *Digital Signal Processing*, vol. 134, p. 103895, 2023.

[5] H. Zhao, T. Wei, W. Zhou, W. Zhang, D. Chen, and N. Yu, Multi-attentional deepfake detection, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Nashville, TN, USA, 2021, pp. 2185–2194.

[6] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, Face X-ray for more general face forgery detection, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2020, pp. 5000–5009.

[7] 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 *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Nashville, TN, USA, 2021, pp. 6454–6463.

[8] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, Thinking in frequency: Face forgery detection by mining frequency-aware clues, in *Proc. 16th European Conf. Computer Vision*, Glasgow, UK, 2020, pp. 86–103.

[9] B. Chen, X. Liu, Z. Xia, and G. Zhao, Privacy-preserving deepfake face image detection, *Digital Signal Processing*, vol. 143, p. 104233, 2023.

[10] B. Liu, B. Liu, M. Ding, T. Zhu, and X. Yu, TI²Net: Temporal identity inconsistency network for deepfake detection, in *Proc. 2023 IEEE/CVF Winter Conf. Applications of Computer Vision*, Waikoloa, HI, USA, 2023, pp. 4680–4689.

[11] R. Caldelli, L. Galteri, I. Amerini, and A. Del Bimbo, Optical flow based CNN for detection of unlearnt deepfake manipulations, *Pattern Recogn. Lett.*, vol. 146, pp. 31–37, 2021.

[12] M. S. Saealal, M. Z. Ibrahim, D. J. Mulvaney, M. I. Shapiai, and N. Fadilah, Using cascade CNN-LSTM-FCNs to identify AI-altered video based on eye state sequence, *PLoS ONE*, vol. 17, no. 12, p. e0278989, 2022.

[13] H. Wang, Z. Liu, and S. Wang, Exploiting complementary dynamic incoherence for deepfake video detection, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 4027–4040, 2023.

[14] Y. Zhu, C. Zhang, J. Gao, J. Gao, X. Sun, Z. Rui, and X. Zhou, High-compressed deepfake video detection with contrastive spatiotemporal distillation, *Neurocomputing*, vol. 565, p. 126872, 2024.

[15] B. Chen, T. Li, and W. Ding, Detecting deepfake videos based on spatiotemporal attention and convolutional LSTM, *Inform. Sci.*, vol. 601, pp. 58–70, 2022.

[16] A. Koteswaramma, M. B. Rao, and G. J. Suma, An intelligent adaptive learning framework for fake video detection using spatiotemporal features, *Signal, Image and Video Processing*, vol. 18, no. 3, pp. 2231–2241, 2024.

[17] J. Wu, Y. Zhu, X. Jiang, Y. Liu, and J. Lin, Local attention and long-distance interaction of rPPG for deepfake detection, *Vis. Comput.*, vol. 40, no. 2, pp. 1083–1094, 2024.

[18] X. Ding, W. Zhu, and D. Zhang, deepfake videos detection via spatiotemporal inconsistency learning and interactive fusion, in *Proc. 19th Annu. IEEE Int. Conf. Sensing,*

Communication, and Networking (*SECON*), Stockholm, Sweden, 2022, pp. 425–433.

[19] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial nets, in *Proc. 27th Int. Conf. Neural Information Processing Systems*, Montreal Canada, 2014, pp. 2672–2680.

[20] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, Face2Face: Real-time face capture and reenactment of RGB videos, in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, NV, USA, 2016, pp. 2387–2395.

[21] J. Thies, M. Zollhöfer, and M. Nießner, Deferred neural rendering: Image synthesis using neural textures, *ACM Trans. Graph.*, vol. 38, no. 4, p. 66, 2019.

[22] R. Tolosana, S. Romero-Tapiador, J. Fierrez, and R. Vera-Rodriguez, deepfakes evolution: Analysis of facial regions and fake detection performance, in *Proc. Int. Conf. Pattern Recognition*, Virtual Event, 2021, pp. 442–456.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929, 2021.

[24] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, L. Wang, and Y. Qiao, UniFormerV2: Unlocking the potential of image ViTs for video understanding, in *Proc. 2023 IEEE/CVF Int. Conf. Computer Vision*, Paris, France, 2023, pp. 1632–1643.

[25] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, RetinaFace: Single-shot multi-level face localisation in the wild, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2020, pp. 5202–5211.

[26] J. Fridrich and J. Kodovsky, Rich models for steganalysis of digital images, *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, pp. 868–882, 2012.

[27] Y. Luo, Y. Zhang, J. Yan, and W. Liu, Generalizing face forgery detection with high-frequency features, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Nashville, TN, USA, 2021, pp. 16312–16321.

[28] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, CBAM: Convolutional block attention module, in *Proc. 15th European Conf. Computer Vision* (*ECCV*), Munich, Germany, 2018, pp. 3–19.

[29] J. Bai, L. Yuan, S. T. Xia, S. Yan, Z. Li, and W. Liu, Improving vision transformers by revisiting high-frequency components, in *Proc. 17th European Conf. Computer Vision* (*ECCV*), Tel Aviv, Israel, 2022, pp. 1–18.

[30] C. Feichtenhofer, X3D: Expanding architectures for efficient video recognition, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2020, pp. 200–210.

[31] K. He, G. Gkioxari, P. Dollár, and R. Girshick, Mask R-CNN, in *Proc. 2017 IEEE Int. Conf. Computer Vision* (*ICCV*), Venice, Italy, 2017, pp. 2980–2988.

[32] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, FaceForensics: A large-scale video dataset for forgery detection in human faces, arXiv preprint arXiv:1803.09179, 2018.

[33] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, Celeb-DF: A large-scale challenging dataset for deepfake forensics, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2020, pp. 3204–3213.

[34] 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.

[35] K. Kim, Y. Kim, S. Cho, J. Seo, J. Nam, K. Lee, S. Kim, and K. Lee, DiffFace: Diffusion-based face swapping with facial guidance, arXiv preprint arXiv:2212.13344, 2022.

[36] S. Zhao, Y. Rao, W. Shi, Z. Liu, J. Zhou, and J. Lu, DiffSwap: High-fidelity and controllable face swapping via 3D-aware masked diffusion, in *Proc. 2023 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, Canada, 2023, pp. 8568–8577.

[37] Z. Hu, H. Xie, L. Yu, X. Gao, Z. Shang, and Y. Zhang, Dynamic-aware federated learning for face forgery video detection, *ACM Trans. Intell. Syst. Technol.*, vol. 13, no. 4, p. 57, 2022.

[38] J. Carreira and A. Zisserman, Quo vadis, Action recognition? A new model and the kinetics dataset, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, HI, USA, 2017, pp. 4724–4733.

[39] 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 *Proc. 2017 IEEE Int. Conf. Computer Vision* (*ICCV*), Venice, Italy, 2017, pp. 618–626.

[40] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, HI, USA, 2017, pp. 1800–1807.

[41] K. Sun, H. Liu, Q. Ye, Y. Gao, J. Liu, L. Shao, and R. Ji, Domain general face forgery detection by learning to weight, in *Proc. 35th AAAI Conf. Artificial Intelligence*, Virtual Event, 2021, pp. 2638–2646.

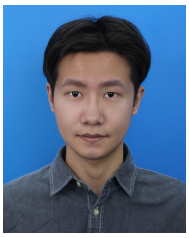
[42] K. Sun, T. Yao, S. Chen, S. Ding, J. Li, and R. Ji, Dual contrastive learning for general face forgery detection, in *Proc. 36th AAAI Conf. Artificial Intelligence*, Virtual Event, 2022, pp. 2316–2324.

[43] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang, End-to-end reconstruction-classification learning for face forgery detection, in *Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, LA, USA, 2022, pp. 4103–4112.

[44] F. Dong, X. Zou, J. Wang, and X. Liu, Contrastive learning-based general deepfake detection with multi-scale RGB frequency clues, *J. King Saud Univ. Comput. Inf. Sci.*, vol. 35, no. 4, pp. 90–99, 2023.

[45] J. Zhang and J. Ni, Domain-invariant feature learning for general face forgery detection, in *Proc. 2023 IEEE Int. Conf. Multimedia and Expo* (*ICME*), Brisbane, Australia, 2023, pp. 2321–2326.

[46] Y. Wang, C. Peng, D. Liu, N. Wang, and X. Gao, Spatial-temporal frequency forgery clue for video forgery detection in VIS and NIR scenario, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 12, pp. 7943–7956, 2023.



Pengfei Yue received the BEng degree in computer science and technology from Nanjing University of Information Science and Technology, China in 2022, where he is currently a master student in electronic information. His research interests include image processing and image forensics.

Beijing Chen received the PhD degree in computer science from Southeast University, China in 2011. He is currently a professor at School of Computer Science, Nanjing University of Information Science and Technology, China. His research interests include digital forensics, image watermarking, color image processing, and pattern recognition.

Zhangjie Fu received the PhD degree from Hunan University, China in 2012. He is currently a professor and the dean at School of Computer Science, Nanjing University of Information Science and Technology, China. He is also filling the post of the director of Engineering Research Center of Digital Forensic affiliated with Ministry of Education, Nanjing University of Information Science and Technology, China. His research interests include digital forensics, blockchain security, artificial intelligence security.