Spatiotemporal Inconsistency Learning and Interactive Fusion for Deepfake Video Detection

DENGYONG ZHANG, Changsha University of Science and Technology, Changsha, China
WENJIE ZHU, Changsha University of Science and Technology, Changsha, China
XIN LIAO, Hunan University, Changsha, China
FEIFAN QI, Changsha University of Science and Technology, Changsha, China
GAOBO YANG, Hunan University, Changsha, China
XIANGLING DING, Hunan University of Science and Technology, Xiangtan, China

With the rise of the metaverse, the rapid advancement of Deepfakes technology has become closely intertwined. Within the metaverse, individuals exist in digital form and engage in interactions, transactions, and communications through virtual avatars. However, the development of Deepfakes technology has led to the proliferation of forged information disseminated under the guise of users' virtual identities, posing significant security risks to the metaverse. Hence, there is an urgent need to research and develop more robust methods for detecting deep forgeries to address these challenges. This paper explores deepfake video detection by leveraging the spatiotemporal inconsistencies generated by deepfake generation techniques, and thereby proposing the interactive spatioTemporal inconsistency learning and interactive fusion (ST-ILIF) detection method, which consists of phase-aware and sequence streams. The spatial inconsistencies exhibited in frames of deepfake videos are primarily attributed to variations in the structural information contained within the phase component of the Fourier domain. To mitigate the issue of overfitting the content information, a phase-aware stream is introduced to learn the spatial inconsistencies from the phase-based reconstructed frames. Additionally, considering that deepfake videos are generated frame-by-frame and lack temporal consistency between frames, a sequence stream is proposed to extract temporal inconsistency features from the spatiotemporal difference information between consecutive frames. Finally, through feature interaction and fusion of the two streams, the representation ability of intermediate and classification features is further enhanced. The proposed method, which was evaluated on four mainstream datasets, outperformed most existing methods, and extensive experimental results demonstrated its effectiveness in identifying deepfake videos. Our source code is available at https://github.com/qff98/Deepfake-Video-Detection

CCS Concepts: • **Computing methodologies** → *Computer vision tasks*; • **Networks** → *Network reliability*.

Additional Key Words and Phrases: Video forensics, Deepfake videos, Spatiotemporal Inconsistency learning, Face recongnition

1 Introduction

Facial forgery detection has become the prime focus multimedia information security. Given that the face serves as a fundamental characteristic of an individual, malicious tampering poses a substantial threat to personal

Authors' Contact Information: Dengyong Zhang, Changsha University of Science and Technology, Changsha, Hunan, China; e-mail: zhdy@csust.edu.cn;Wenjie Zhu, Changsha University of Science and Technology, Changsha, Hunan, China; e-mail: wenjiezhu@stu.csust.edu.cn; Xin Liao (Correspondence author), Hunan University, Changsha, Hunan, China; email: xinliao@hnu.edu.cn; Feifan Qi, Changsha University of Science and Technology, Changsha, Hunan, China; e-mail: qifeifan1998@163.com; Gaobo Yang, Hunan University, Changsha, Hunan, China; e-mail: yanggaobo@hnu.edu.cn; Xiangling Ding, Hunan University of Science and Technology, Xiangtan, China; e-mail: xianglingding@163.com.

ACM 1551-6865/2024/9-ART
https://doi.org/10.1145/3664654

privacy. With the increasing application of face recognition in daily life, there is related research in the field of face recognition dedicated to improving face detection performance [55], such as utilizing the correlation of multi-view information [29] in images for detection, which increases the awareness of the social hazards brought about by deepfake. Moreover, the emergence of the metaverse and the integration of deepfake technology with it have made the issue of deepfakery more complex and harder to detect. Within the metaverse, we are all represented by avatars, yet the sophistication of deepfake technology makes it alarmingly easy to mimic someone's avatar, thereby serving the purposes of the impersonator. Founders of the Metaverse [6] have also become aware of the potential risks deepfake pose to their platforms. While some recent legislation has been enacted to address the issues arising from deepfakery, there isn't yet a dependable legal solution within the metaverse. Consequently, developing effective means to detect deepfake using pertinent technologies is an urgent matter for ensuring the security of the metaverse. On the surface web, the misuse of deepfake videos has led to a series of issues. Similarly, in the metaverse, the presence of deepfake videos may erode trust, damage reputations, invade privacy, and potentially submerge the entire metaverse in misleading information. The task of detecting deepfake videos can be categorized into two distinct approaches: frame and video-level methods. Within the frame-level approach, certain techniques aim to capture indications of tampering within the spatial domain, as exemplified in prior studies such as [2, 13, 54, 63]. Conversely, other studies have sought to leverage frequency information to improve overall performance, as demonstrated in [18, 26, 43, 61]. Although the frame-level method obtains better in-dataset detection performance, it ignores the inter-frame temporal features of the video; therefore, it needs to be improved for cross-dataset detection. The generation of deepfake videos typically involves the creation of individual frames using generative adversarial network GAN-based models. Consequently, temporal information is often lacking between consecutive frames, leading to a noticeable lack of smoothness in the resulting deepfake video compared with the original footage. To address this issue, researchers have leveraged techniques to incorporate temporal information. However, in methods based on 3D convolutional neural network (3DCNN) [27], the 3D convolution for feature extraction imposes a significant computational burden. In addition, methods based on recurrent neural networks (RNN) [33] may not be able to fully learn spatially inconsistent features or extract spatiotemporal features simultaneously, leading to suboptimal performance. Notably, none of these methods effectively utilizes multiple feature representations of the image information.

As widely recognized, the production of a deepfake video involves the generation of a fake target face and its subsequent fusion with the background using a distinct model [1]. However, this operation may result in visual artifacts at the edges of the synthesized faces, leading to inconsistencies in spatial structures. Notably, the structural information and fundamental spatial relationships of the original image are preserved in the phase component of the Fourier frequency domain [57].

Because it is not possible to distinguish real and fake images directly from visualizations of the phase spectrum, processing is required to obtain image features that can be learned by CNNs. Therefore, this study proposes a phase-based frame reconstruction method to enhance the structural information of the image, highlighting the spatial inconsistencies between real and fake images. Subsequently, the phase-aware stream of spatiotemporal inconsistency learning and interactive fusion (ST-ILIF) is utilized to capture the structural inconsistency information present in the phase-based reconstructed frames.

Spatiotemporal inconsistency information plays a significant role in the detection of deepfake videos. Therefore, we designed a temporal difference module (TDM) to capture spatiotemporal inconsistency features from the difference information between consecutive frames, and placed the TDM at the input of the sequence stream of ST-ILIF. As shown in Fig. 1, there is no apparent inconsistency between the original and deepfake videos in the image spatial domain. Further-more, the frame difference images in the spatial domain show that deepfake video sequences have greater inter-frame differences than real video sequences. This suggests that the spatiotemporal inconsistency may be a helpful cue for deepfake video detection. We also present the frame-difference image

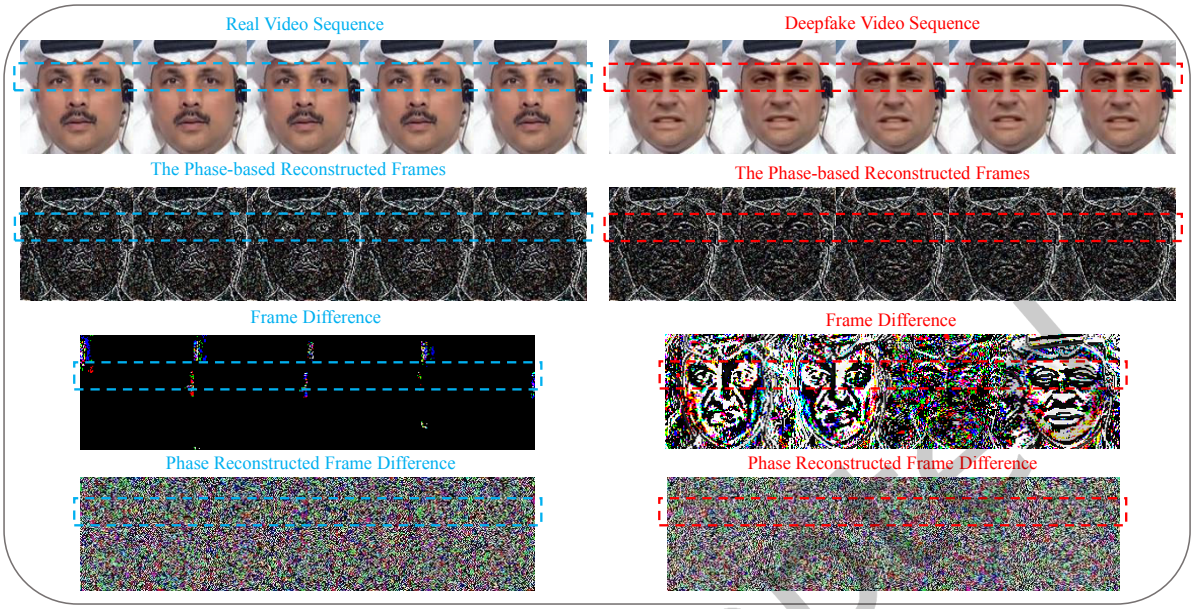

Fig. 1. Example of frame difference image and phase reconstruction image of real video and Deepfake video.

of the phase-based reconstructed frame. No inconsistent information is observed, indicating that the frame difference operation does not apply to the phase-based reconstructed frame.

Given that ST-ILIF operates in a two-stream manner in, which each stream extracts inconsistent features from the spatial, temporal, and frequency domains, a channel-attention-based feature interaction module is proposed to facilitate complementary feature learning in the intermediate layers of the two streams. Furthermore, inspired by [32], an adaptive feature fusion mModule (AFFM) is proposed to adaptively fuse two-stream features and enhance the repre-sentational capacity of classification features. As an extension of our previous work [7], this study will further learns tampered traces from the spatiotemporal and frequency domains to detect deepfake videos. The main contributions of this study are as follows:

- The present study defines deepfake video detection as a process for learning spatiotemporal inconsistencies and fusing multiple features. To achieve this objective, a two-stream model called ST-ILIF, is proposed for video-level deepfake detection. This network efficiently captures spatial and temporal tampering traces by leveraging a more complementary representation.
- A new approach is proposed for extracting structural features from images, whereby Fourier-domain phase components are used to reconstruct the frames, emphasizing spatially inconsistent features. In addition, a temporal difference module is introduced to captures spatiotemporal in-consistency features from the difference information between consecutive frames. The TDM, along with the phase recon-struction frames, enables the effective capture of spatiotemporal inconsistency information in deepfake videos. To promote efficient interaction and integration of features from both streams, we propose two additional modules: the channel attention-based feature interaction module (CAFIM) and adaptive feature fusion module, namely AFFM, which interact and fuse the intermediate features, to discriminate the features of the phase-aware and sequence streams, respectively. The proposed framework, ST-ILIF, is designed for video-level

deepfake detection and efficiently captures spatial and temporal tampering traces in a more complementary representation.
- Extensive experimental results showed that TDM can accurately capture temporal difference information, whereas CAFIM and AFFM can effectively facilitate two-stream feature interaction and fusion. With the addition of these modules, the proposed network outperformeds most detection methods on public datasets.

The remainder of this paper is organized as follows. In Section II, a summary of related research is provided. In Section III, deepfake videos are examined from the perspective of video quality. Section IV presents the proposed video-level detection approach. In Section V, the implementation of the proposed method is outlined, and the experimental results are presented. Finally, conclusions are drawn in Section VI.

2 Related Work

This section introduces related work on deepfake detection, including deepfake generation, video- tampering forensics, and deepfake video forensics.

2.1 Deepfake Generation

Previous face forgery video generation algorithms were based on computer graphics. Face2Face [48] is a real-time face reenactment technique that aims to transfer the facial expressions from a source face to a target face, while preserving the identity of the target individual. This method utilizes advanced computer vision algorithms to achieve a high degree of fidelity in the re-enacted facial expressions. FaceSwap [44] is a face manipulation technique that aims to replaces the appearance of the source face with that of the target face. This method utilizes facial landmarks to extract relevant facial regions, that are subsequently used to construct a 3D model. The generated model is refined by minimizing the difference between the projected shape and local landmarks, using texture information from the input image. Finally, the model is back-projected onto the target image and color-corrected. Traditional forgery methods often have low synthesis quality and slow execution speed. In contrast, existing face-swapping techniques based on deep learning have significantly improved synthesis quality and runtime speed. The deepfake [1] technique employs an encoder and two decoders for facial manipulation , whereby the source face is replaced by the target face. The encoder is trained to learn the common attributes of the two faces, whereas the decoders reconstruct the original image by leveraging the hidden features generated by the encoder. Some of these methods are based on generative adversarial networks. Face swapping GAN (FSGAN) [37] is capable of face swapping and multi-view face interpolation for any two face images and achieves a good face swapping effect. Liu[28] et al. proposed two novel regularizations to enhance the 3D accuracy of face generation models and the 2D image quality. DeepFaceLab [41] is deep learning-based facial recognition and face-swapping software that can achieve high-precision face detection, key point localization, feature extraction, and expression synthesis. It can achieve highly realistic face-swapping effects. The core concept of the algorithm is to extract and reconstruct facial features using neural networks, followed by synthesis and transformation operations.

2.2 Deepfake Videos Forensics

In recent years, the rapid spread of deepfake videos on the Internet has disrupted audio-visual perceptions, prompting the development of various frame-level counter-measures for their detection. For example, Li and Lyu [22] observed face-warping artifacts in deepfake videos and devel-oped a detection approach based on these observations. FRLM [35] improves model generalization performance by learning forgery region awareness and ID-independent features. Zhao [64] distinguish real and fake faces by detecting whether the internal and external features of a face belong to the same ID. However, these frame-level detectors neglect the temporal characteristics

of videos, which limits identification accuracy. Therefore, it is crucial to develop new approaches that consider the spatiotemporal consistency of videos for deepfake detection.

Video-level deepfake detection approaches often prioritize the extraction of temporal features to optimize the model performance in detecting manipulated videos. A multi-rate excitation network [40] was proposed to extract long- and short-term spatiotemporal inconsistency information from forged videos. Wang [51] proposed a complementary dynamic interaction network for deepfake video detection using multi-task learning, thereby achieving better generalization capability. Hu et al.[16] proposed a dual-stream network that utilizes frame-level and temporal-level features to jointly detect compressed videos. Ge et al. [10] found temporal inconsistencies in facial components across frames in deepfake videos, leading to the proposal of a predictive detection method termed 'Latent Pattern Sensing'. Additionally, they improved the detection performance of the model by incorporating an attention mechanism. In contrast to the aforementioned approaches, our study focuses on multi-frame difference motion modeling for deepfake video detection. The crucial aspect lies in capturing spatiotemporal inconsistencies. Moreover, enhancing the interaction and fusion between the phase-aware and sequence streams is beneficial for learning more effective video-level representations.

3 Motivation and Feasibility Analysis

3.1 Unreferenced analysis of deepfake videos

This section, presents a quantitative analysis of the deepfake video using the no-reference video evaluation method.

As a deepfake video is generated frame by frame, it lacks temporal continuity, which is an excellent clue for detecting deepfake videos. In addition, after a series of post-processing steps, such as video compression, the gap between real and fake videos becomes more significant. Video quality assessment, especially no-reference video quality assessment, can be used to evaluate videos without the original information. We chose VSFA to evaluate videos in the FaceForensics++ training set. Specifically, VSFA [21] is a no-reference video evaluation method that reliably evaluate wild videos using content-aware features and temporal memory modeling. However, because the forged videos only tamper with the face, evaluation of the complete video is not informative; Therefore, we synthesized videos out of the extracted face image sets and evaluated these videos using the original weight files provided by VSFA. The results presented in Table 1, report the average evaluation scores of 720 videos in the training set.

Table 1. Video quality evaluation results of the FaceForensics++ training set. The results report the average evaluation scores of all videos.

	Real	Deepfakes	Face2Face	FaceSwap	NeuralTextures
c23	0.9211	0.9000	0.9154	0.9150	0.9170
c40	0.8532	0.8347	0.8485	0.8449	0.8522

In Table 1, the scores of the fake videos are all lower than those of the real videos. Among the lightly compressed videos (c23), the evaluation scores of the videos generated by the deepfake tampering method differ the most from those of real videos, indicating that the quality of the fake videos generated by Deepfakes is the worst. Neural Textures is one of the more advanced forgery methods. Using neural texture methods can significantly improve synthesis quality, such that the method generates the smallest gap between the forged and real videos. The same phenomenon is observed in heavily compressed videos (c40). The evaluation results of the non-referenced videos show that there is a difference in quality between real and fake videos. This difference is manifested both in the

spatial domain inconsistency of the real and fake videos implying that the forged videos lack some temporal continuity and the changes between consecutive frames are not smooth enough, thereby resulting in lower scores in the evaluation results. Accordingly, we model the spatiotemporal inconsistency information of real and fake videos for deepfake video detection.

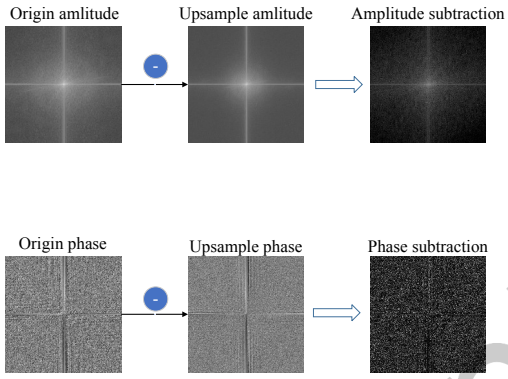

Fig. 2. Schematic diagram of frequency domain analysis. Divide the frequency domain into phase spectrum and amplitude spectrum for analysis. The first line is the residual of the amplitude spectrum of the original image and the upsampled amplitude spectrum. The second line is the residual between the phase spectrum of the original image and the upsampled phase spectrum.

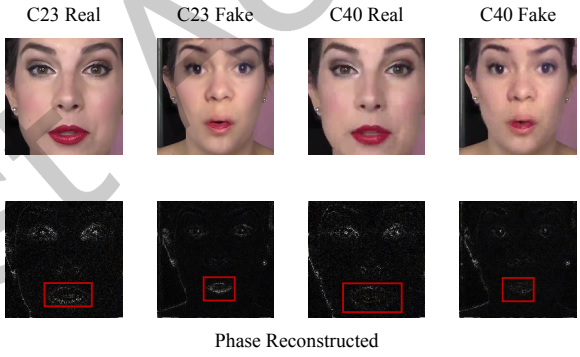

Fig. 3. Frequency domain interpretation. Phase reconstruction using different compressed versions of images. It can be seen that under different degrees of compression, the inner and outer lips of the real face still have obvious inner and outer edge structures, while the lips of the fake face are surrounded by mixed blur artifacts.

3.2 Frequency domain analysis

The feasibility of our method, which uses only the phase spectrum in the frequency domain for the analysis, was confirmed through the following experimental analysis. The proposed method divides the frequency domain into

amplitude and phase spectra, and analyzes the correlation between the two spectra and the forgery method. First, most general-purpose forgery methods inevitably require upsampling operations. Therefore, we first determined the upsampling operation that would have a greater impact on the amplitude or phase spectrum in the frequency domain. The visualization results are presented in Fig. 2. After binarizing the residual image, the absolute value is taken. The brighter the color, the more obvious is the difference. The average pixel difference of the phase spectrum is significantly larger than that of the amplitude spectrum, indicating that the phase spectrum is more sensitive to the frequency-domain upsampling operation. To analyze the robustness of the method, we performed phase reconstruction on the images under different compression levels, which is visualized in Fig. 3. Under different degrees of compression, the real image displays clear edge information and structure of the lips, whereas the fake image is mixed with artifact information and does not delineate the lip structure well. This indicates that the phase spectrum is very robust to compression. Based on these two points, the phase spectrum in the frequency domain was used to detect deepfake forgeries.

The proposed method is not the first article to analyze Deepfake from the frequency domain. There have been some previous articles that analyzed from the frequency domain, but the proposed method is very different from them. Their methods directly change the image to the frequency domain, then directly use a fixed low-pass filter for filtering, and finally inversely transform it back to the spatial domain to obtain image artifacts. We know that the frequency domain is susceptible to compression, but their method does not take this into account, which causes the filtered results to easily appear ringing, affecting the detection results. The proposed method does not consider the amplitude spectrum in the frequency domain but starts with the phase spectrum because the amplitude spectrum is easily affected by compression in the frequency domain but the phase spectrum is not easily affected. At the same time, the phase spectrum is more susceptible to the influence of upsampling operations, and there must be upsampling operations during the forgery process. So we only used the phase spectrum in the frequency domain for Deepfake detection, which was not mentioned in the previous methods.

4 ST-ILIF

This section provides a comprehensive presentation of the proposed approach, namely ST-ILIF.

4.1 Overview

The problem of identifying deepfake videos at the video-level can be formulated as a binary classification task. Specifically, given a video sequence of dimensions $T \times C \times H \times W$, where T, C, H, and W denote the number of frames, number of channels, height, and width of each frame, respectively, a model design is required to output a probability that indicates whether the video sequence is fake or real. This task is formulated as a spatiotemporal inconsistency learning process that involves the interactive fusion of multiple features, which are subsequently integrated into the proposed phase-aware and sequence streams. The architecture of the model is illustrated in Fig. 4. The proposed network operates in a two-stream manner, comprising a phase-aware stream and a sequence stream. The input to the phase-aware stream is a phase-based reconstructed frame, that leverages the property that the phase component in the frequency domain representation of the Fourier-transformed image carries structural information, whereby the structural features of the image are reconstructed solely based on the phase component, to expose traces of tampering and improve model generalization performance. The reconstructed frames are concatenated onto the original frames and fed into a backbone network specifically designed for this type of data for feature extraction. The generation of deepfake videos typically overlooks inter-frame coherence, which highlights the need to train the sequence stream on time-varying attributes to capture the distinctions between consecutive multi-frames, for short-term motion modeling. To address this issue, a temporal difference module (TDM) was integrated with convolutional gated recurrent units (ConvGRU) within a shallow network architecture to capture spatiotemporal incoherence. For feature extraction, we employed

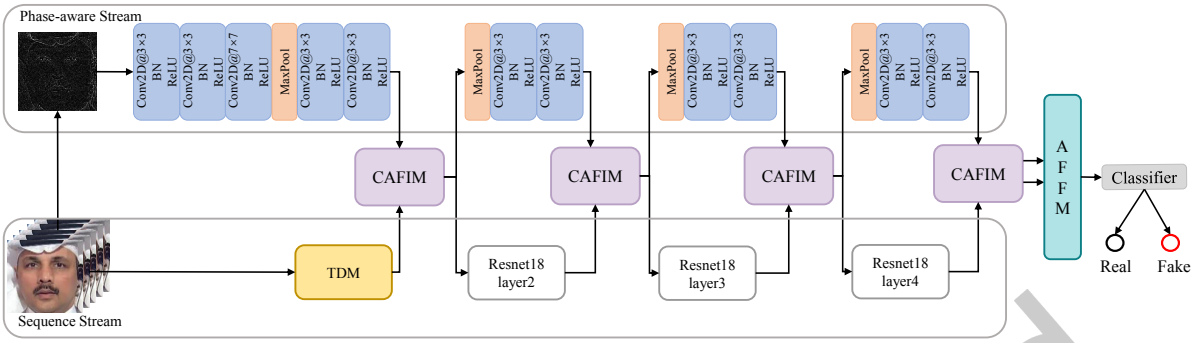

Fig. 4. Example diagram of the ST-ILIF network architecture, which consists of phase-aware and sequence streams. The former employs phase-based reconstructed frames to model and capture spatial structural inconsistencies, while the latter utilizes the TDM to learn spatiotemporal motion inconsistencies. The spatiotemporal inconsistency is subsequently interactively fused with the CAFIM and the AFFM.

the last three layers of the residual network Resnet18 [15]. Furthermore, a channel attention-based feature interaction module (CAFIM) was proposed to facilitate the interplay between frame and sequence pathways, thereby augmenting their complementary characteristics. Finally, an adaptive feature fusion module (AFFM) was constructed to adaptively fuse the two stream features, to aid in video-level decisions.

The aim was to make the network learn the maximum difference between consecutive frames of a fake video and the minimum difference between consecutive frames of a real video. Therefore, windows that are too long may result in large differences between the first and last frames, making the feature distributions of the real and fake videos similar over long periods, resulting in the network not being able to learn discriminative features. Therefore, five consecutive frames were used as input to the network.

4.2 Phase-aware Stream

Deepfake video generation is typically accompanied by visually distinguishable artifacts, including a checkerboard pattern resulting from up-sampling operations and blending edges between the real and tampered regions within each frame. Collectively, these artifacts are referred to as spatial structural inconsistency and can be easily identified [11]. However, [14, 38, 39, 42, 58] the phase component has the property of semantic preservation and highlighting the phase information can enhance the generalization performance of the model. To avoid overfitting the spatial information of the original image in the CNN model, which could potentially impair generalization performance, we introduced a phase-based frame reconstruction method and a CNN model specialized in extracting features from this type of data. A detailed of the operational process of the phase-aware stream is as follows.

Specificlly, let $f(x, y)$ denote a grayscale image with dimensions $H \times W$. The Fourier transform of $f(x, y)$ can be expressed as: where i are Euler's number and imaginary units, respectively. The function $F(u, v)$ can be alternatively represented in terms of its magnitude $A(u, v)$ and angle $P(u, v)$.

$$\begin{aligned} F(u,v) &= A(u,v)e^{-iP(u,v)} \\ A(u,v) &= \left[R^2(u,v) + I^2(u,v)\right]^{1/2} \\ P(u,v) &= \arctan\left[\frac{I(u,v)}{R(u,v)}\right] \end{aligned} \tag{1}$$

where $R(u,v)$ and $I(u,v)$ represent the real and imaginary parts of $F(u,v)$, respectively.

The phase components of the video frames contain crucial structural-semantic information. Nevertheless, in deepfake videos, discerning the phase components from those of real videos is exceedingly challenging. Therefore, to retain the spatial structural information of the source image, we removed the amplitude component and performd a Fourier inverse transformation based solely on the phase component to reconstruct the original image. This is accomplished by assigning a fixed constant value to $A(u,v)$, which yields the following phase-based reconstruction expression for the frame:

$$f(x,y) = |F^{-1}\left(const \times e^{-iP(u,v)}\right)| \tag{2}$$

where F^{-1} denotes the inverse Fourier transformation.

$$F(u,v) = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} f(x,y)e^{-i2\pi\left(\frac{ux}{H}+\frac{vy}{W}\right)} \tag{3}$$

To demonstrate the efficacy of the phase-based reconstruction frame, we compared it with the amplitude-

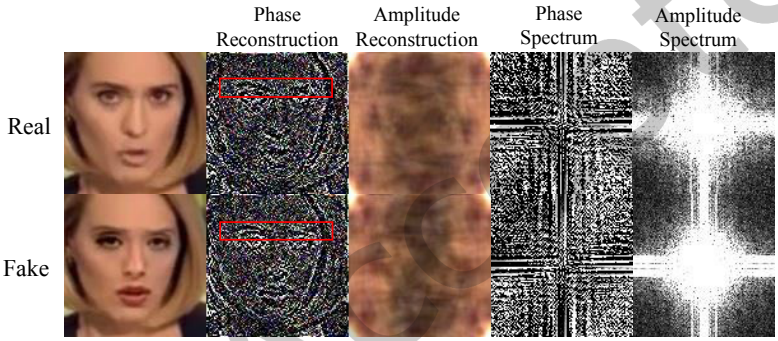

Fig. 5. The results of comparing the reconstructed frames based on their phase and amplitude, and the phase and amplitude spectra, are presented.

based reconstruction frame and examined the respective phase and amplitude spectra, as shown in Fig. 5. It is evident that the phase-based reconstructed frames capture facial structural information remarkably well, particularly with respect to the inconsistent information present in the eyebrows. Conversely, the amplitude-based reconstructed frames fail to preserve useful spatial information. Additionally, neither the amplitude nor the phase spectra contain any meaningful inconsistent information, which renders them unsuitable for training the model to discern deepfake videos. Consequently, the optimal approach is to reconstruct the phase information to emphasize structural features.

A single-channel grayscale image was obtained by averaging the input of five consecutive frames in the channel dimension and phase reconstruction was performed on this grayscale image. The phase-reconstructed image was cascaded with the grayscale image as the input to the phase-aware stream, and feature extraction was performed. Some edge detection operators can achieve results similar ti those of the phase-based frame reconstruction method. The classical edge detection operators Sobel and Prewitt were seleced for the comparison. The kernel parameters of these two operators are shown in Fig. 6. In addition, ablation experiments were performed to demonstrate the effectiveness of the phase-based frame reconstruction method.

The network structure of the phase-aware stream is designed specifically for feature extraction of phase-reconstructed images. Initially, the input data go through two convolutional layers, each with a stride of 1 and a

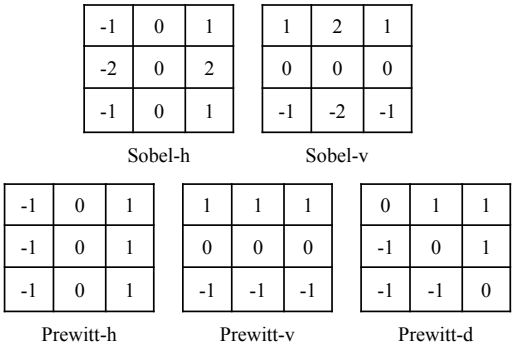

Fig. 6. *prewitt* and *sobel* operators.

kernel size of 3. This process enhances the input features by increasing the number of input channels to 64 while maintaining the original resolution. The feature map is subsequently downsampled via a convolutional layer with a stride of 2 and a kernel size of 7. Further downscaling of the feature map is performed using MaxPool, followed by additional refinement through two convolutional layers, each with a stride of 1 and a kernel size of 3. The feature map size at this stage is equivalent to that of the TDM module output of the sequence stream. Two sets of features from the phase-aware and sequence streams aare fed in to the first CAFIM module for feature interaction. The phase-aware stream also includes three sub-modules, each performing downsampling using MaxPool, and two convolutional layers for feature extraction. The output features of each submodule interact with the features of the sequence stream. The final feature fusion is performed using AFFM.

4.3 Sequence Stream

In view of the fact that deepfake videos are typically manipulated frame by frame, the lack of temporal coherence between consecutive frames represents a crucial cue for deepfake videos detection. To fully leverage this feature, we introduced the temporal difference module, that utilizes 2D convolutional networks to capture spatiotemporal inconsistency information, akin to [52]. The inconsistent temporal features in Deepfake videos are inconspicuous, and may gradually disappear as the network deepens. To address this issue, we employed a convolutional gated recurrent unit (ConvGRU) to refine the spatiotemporal features and enhance the representational capacity of local spatiotemporal features. Fig. 7 illustrates the overall structure. Specifically, given a sequence of input video frames $[I_{t-2}, I_{t-1}, I_t, I_{t+1}, I_{t+2}]$, the differences between consecutive frames in the temporal domain are computed and concatenated. The deepfake video detection task only considers whether forgery in the face region; therefore, only the face region needs to be cropped for model training and classification. However, facial images of consecutive frames exhibit slight variations, and there is more information redundancy. To reduce the computational complexity, we used average pooling to downsample prior to refining the feature representation using a convolutional layer with a stride of 2 and a kernel size of 7×7. Subsequently, the feature representation of the spatiotemporal information was further refined and enhanced by utilizing ConvGRU. The ConvGRU architecture is shown illustrated in Fig. 8.

Two pathways are involved in TDM. The first pathway extracts features from the middle frame of the input sequence, I_t, through a series of convolutional and max-pooling layers to generate feature map X_t. The second pathway concatenates the inter-frame differences of the input frames to obtain spatiotemporal difference features with are then downsampled using an average pooling layer to mitigate information redundancy and further processed with convolutional and max-pooling layers to extract features, The downsample feature map results

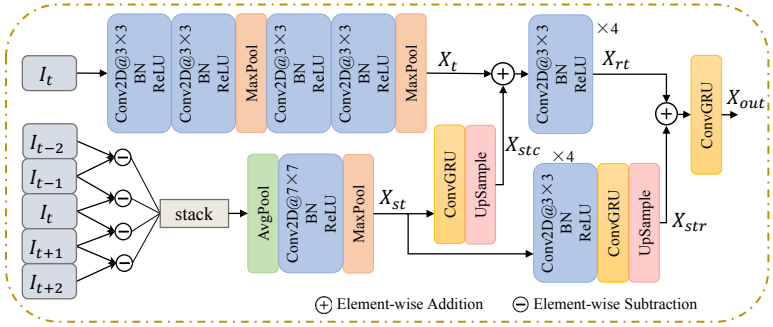

Fig. 7. The Temporal Difference Module (TDM) is designed to extract spatiotemporal difference features from consecutive frames. The spatiotemporal features obtained are then subjected to further refinement via the utilization of ConvGRU.

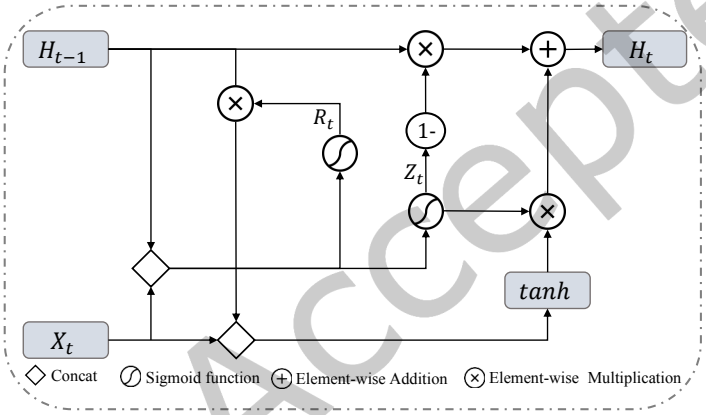

Fig. 8. ConvGRU block with feature map as input.

in the feature representation X_{st}. These features are refined through the first ConvGRU unit to capture spatiotemporal features, and the upsampled feature map X_{stc} is added to the middle frame feature map X_t, allowing the spatiotemporal differences to be perceived in a single frame. The features are further refined using four convolutional layers, resulting in the feature map X_{rt}; X_{st} is then passed through another four convolutional layers to extract features, followed by a second ConvGRU unit to enhance feature representation. After obtaining the upsampled feature map X_{str}, X_{rt} and X_{str} are added before inputting into the third ConvGRU unit to obtain the output feature map X_{out} of the temporal difference module. We implemented the method outlined in [25] to process the input features within the ConvGRU unit. Specifically, the input features are split along the channel dimension, with half of the features being fed into the ConvGRU unit to learn temporal motion information and the other half being retained to preserve spatial features. Finally, the two sets of features are concatenated to obtain the output results. The detailed steps of this process are illustrated in Fig. 9.

4.4 Channel Attention-based Feature Interaction Module

The phase-aware stream primarily captures structural inconsistencies, whereas the sequence stream emphasizes temporal inconsistencies. Both inconsistency features can be used as favorable cues for Deepfake video detection.

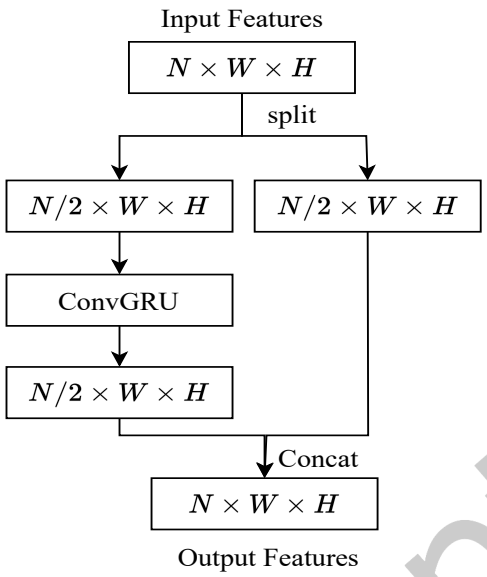

Fig. 9. ConvGRU data processing flow.

Therefore certain methods can be used to make these two features complement each other to facilitate feature representation in the middle layer of the network, thus effectively preventing inconsistent information from disappearing gradually as the network deepens. A channel attention-based feature interaction Module (CAFIM) was designed for this purpose. The structure of the CAFIM module is illustrated in Fig. 10.

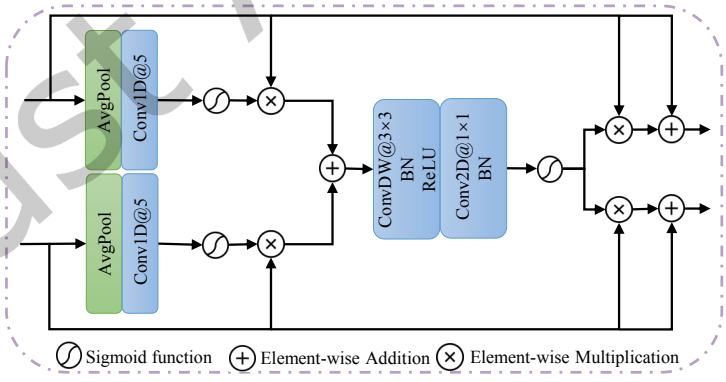

Fig. 10. CAFIM module structure diagram. We use channel attention to obtain the importance of the two sets of features.

To evaluate the relevance of distinct channels, feature representations from the phase-aware stream (X_f) and the sequence stream (X_s) are subjected to channel attention weighting [53], to determine the relative importance

of each channel within the feature representation. Then the aggregated features X_{agg} can be represented as:

$$\begin{aligned} X_{agg} = X_f \times \sigma(Conv1D(AP(X_f))) \\ +X_s \times \sigma(Conv1D(AP(X_s))) \end{aligned} \tag{4}$$

where $Conv1D$ has a stride of 1 and a kernel size of 5; AP denotes average pooling. Subsequently, the features are refined through a deep-wise convolution operation applied independently to each channel, and a regular convolution operation is further utilized to integrate the information across different channels. This process is expressed as follows:

$$X_w = \sigma(BN(Conv2D(Re(BN(ConvDW(X_{agg})))))) \tag{5}$$

where the $\sigma, BN, Re, ConvDW$, and $Conv2D$ operations correspond to the sigmoid activation function, batch normalization, ReLU activation function, 3×3 deep-wise convolution, and 1×1 regular convolution, respectively. Finally, attention maps are separately added to the original features to enhance the maps and avoid gradient disappearance. The resulting refined features are denoted as X_r.

$$\begin{aligned} X_{rf} = X_f + (X_f \times X_w) \\ X_{rs} = X_s + (X_s \times X_w) \end{aligned} \tag{6}$$

4.5 Adaptive Feature Fusion Module

Given that the prediction outcomes at the video level are determined by two distinct streams, it is necessary to employ specialized techniques to effectively integrate the features of these two streams to reach a final decision. The conventional fusion methods include summation and cascading. However, a simple summation or cascading cannot fully utilize the features of the two streams to effectively discover the important information in the feature map in a global-local manner. Therefore, we propose the adaptive feature fusion module to adaptively select discriminative features from a global-local perspective, thus enhancing the feature representation capability for prediction. Fig. 11 illustrates the organization of the AFFM module.

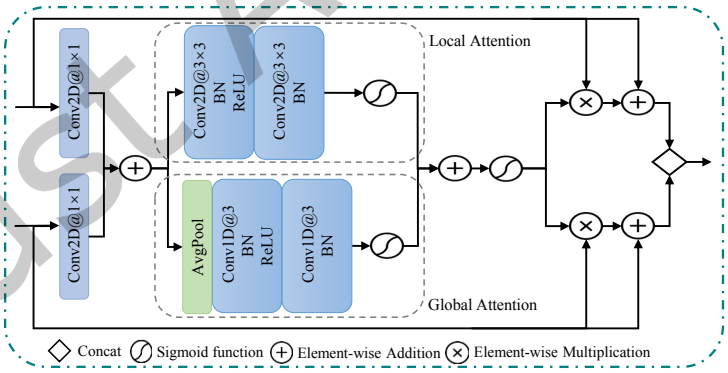

Fig. 11. AFFM adaptively fuses two-branch features from a global-local manner to improve the representation of classification features.

Let X_{ff} and X_{fs} denote the feature maps generated by two streams. These features are first element-wise added, denoted as:

$$X_{add} = Conv2D(X_{ff}) + Conv2D(X_{fs}) \tag{7}$$

Subsequently, for the local attention, two convolutional layers with a 3×3 kernel size are employed and expressed as follows:

$$L_{att} = \sigma(BN(Conv2D(Re(BN(Conv2D(X_{add})))))) \tag{8}$$

To perform global attention, we first compute the global representation of each channel via average pooling. Subsequently, we employ two $Conv1D$ convolutions with a kernel size of 3 to aggregate the features across different channels, defined as:

$$G_{att} = \sigma(BN(Conv1D(Re(BN(Conv1D(AP(X_{add}))))))) \tag{9}$$

The fusion process of the obtained local attention map L_{att} and global attention map G_{att} is performed as follows:

$$Fuse_{att} = \sigma(L_{att} + G_{att}) \tag{10}$$

The resulting global-local attention map is denoted as $Fuse_{att}$, which combines the local and global attention maps. The ultimate features that are utilized for classification are computed as:

$$\begin{aligned} X_{final} = Concat((X_{ff} + (X_{ff} \times Fuse_{att})), \\ (X_{fs} + (X_{fs} \times Fuse_{att}))) \end{aligned} \tag{11}$$

After the feature concatenation operation (denote as $Concat$), the resulting feature map X_{final} is fed into the last layer of the proposed network to obtain the video-level classification results.

5 Experiments

This section expounds on the implementation of the ST-ILIF technique, including the experimental configuration and assessment of its performance relative to state-of-the-art methodologies.

5.1 Evaluation Datasets

FaceForensics++ (FF++) [44]: The FF++ dataset is widely employed in facial forgery detection research, comprising 4,000 manipulated videos generated by Deepfakes(DF), Face2Face(F2F), FaceSwap(FS), and NeuralTextures(NT). To ensure a fair comparison, subsets of the training, validation, and test sets were constructed using the official video list with 720, 140, and 140 videos, respectively. The extract video frames were extracted by adhering to the convention of most existing studies, whereby the first 270, 140, and 140 frames were extracted for the training, validation, and test sets, respectively.

Celeb-DF [23]: The Celeb-DF dataset comprises 5639 high-quality deepfake videos with enhanced facial details. In line with the standard practice of current detection methodologies, it was leveraged as a benchmark dataset to evaluate the generalization capability of the proposed model.

DeeperForensics-1.0 (DF-1.0) [19]: DF-1.0 is a large-scale dataset that provides real-world scenarios for face forgery detection. The dataset is composed of 1000 original videos sourced from FF++ and recorded by 100 actors. The end-to-end densely fused variational autoencoder (DF-VAE) was used to swap each face of the 100 identities, distributing into 10 distinct target videos. To emulate real-world conditions, the dataset introduces various perturbations into the videos. In this study, the "std/rand" subset of the dataset was utilized to execute the generalization tests on the provided test set list.

Deepfake Detection Challenge dataset (DFDC) [8]: DFDC is a large dataset commonly used in the field of deepfake detection. It contains nearly 400G of forged videos generated using different tampering methods. Among all videos, we randomly selected 2000 real videos and the corresponding 5843 forged videos for model generalization performance evaluation.

5.2 Implementation Details

In this study, the proposed ST-ILIF method was implemented using the PyTorch framework and trained on an NVIDIA 3060Ti GPU. The ST-ILIF input sequence size was established as $5 \times 3 \times 224 \times 224$, and the amplitude *const* in Eq. (3) was fixed at 25000. The network was trained end-to-end using the Adam optimizer [20]. Binary cross-entropy, which is widely employed in the realm of deepfake detection, was adopted as the loss function, as shown in Eq. (12). where, y_i denotes the label of a given sample i and p_i represents the likelihood of sample i being predicted as a positive class. Furthermore, the cosine annealing learning rate decay was integrated, to gradually decrease the initial learning rate from 0.0005 to 0.0001. The model was trained for 50 epochs with a batch size of 32. To integrate the temporal information, random batch-level flipping was implemented as a data augmentation technique. To obtain the final detection score at the video level, all the sequences within each video were processed in batches, thereby computing the mean of the predicted scores.

$$L = \frac{1}{N} \sum_{i} -[y_i log(p_i) + (1 - y_i) log(1 - p_i)] \tag{12}$$

5.3 Ablation Study

A series of experiments were evised to systematically demonstrate the efficacy of each ST-ILIF module with respect to the video-level accuracy metrics on the c23 subset of the FF++ dataset.

5.3.1 Phase-aware stream Evaluation. The performance comparison of various edge detection operators and reconstruction methods is presented in Table 2. In amplitude reconstruction method the frames are reconstructed using only the amplitude component. The results reveal that our proposed phase reconstruction method outperforms other methods, which can be attributed to the phase-aware network structure design. The phase reconstruction approach offers an additional perspective for extracting image structural information and is better suited for face tampering detection tasks.

Table 2. Comparison of edge detection operators with different reconstruction methods.Considering that the discriminative ability of the model towards positive and negative samples more intuitively reflects the evaluation results and ensures objectivity and comprehensiveness, the evaluation metric adopts AUC.

Architecture	Methods	FF++ c23
Phase-aware stream	Sobel	95.29%
	Prewitt	95.71%
	Amplitude Reconstruction	95.14%
	Phase Reconstruction	**96.00%**

5.3.2 sequence stream Evaluation. The performance of the sequence stream is presented in Table 3. Notably ResNet18, which was selected as the backbone network in the sequence stream, also reported video-level accuracy. However, the performance of ResNet18 is relatively poor because of its inability to effectively utilize temporal information in videos, On the other hand, incorporating TDM and stacked ConvGRU units in the sequence stream proved to be effective in refining the temporal information, resulting in an improvement of 3.57% over the baseline.

5.3.3 Two-stream Evaluation. Table 4 reports the performances of the different modules in a two-stream structure. The default feature fusion method is a concatenated structure. A two-stream network can extract features from

Table 3. Sequence stream ablation experiments.The evaluation metric is AUC.

Architecture	Models	FF++ c23
sequence stream	ResNet18	92.14%
	TDM+ResNet18	**95.71%**

various perspectives. The design of the CAFIM module effectively facilitated two-stream feature interaction, thus improving the performance to 96.14%. AFFM can perform feature fusion in a global-local manner, with a slight degradation in performance when feature interaction is not performed. When the CAFIM and AFFM modules are combined, the two-stream network achieves the best performance of 97.29%, demonstrating the ability of the designed modules to mutually facilitate the detection performance of the model.

Table 4. Performance evaluation of different modules with two-stream structure.The evaluation metric is AUC.

Architecture	CAFIM	AFFM	FF++ c23
Two-stream	×	×	95.57%
	✓	×	96.14%
	×	✓	96.57%
	✓	✓	**97.29%**

Table 5. The evaluation of different methods on FF++ at various compression levels is presented in the current study. The performance comparison with other SOTA methods is carried out based on the results reported in [11].The evaluation metric is AUC.

	FF++ c23				FF++ c40			
Methods	DF	F2F	FS	NT	DF	F2F	FS	NT
C3D [49]	92.86%	88.57%	91.97%	89.64%	89.29%	82.86%	87.86%	87.14%
I3D [5]	92.86%	92.86%	96.43%	90.36%	91.07%	86.43%	91.43%	78.57%
FaceNetLSTM [46]	89.00%	87.00%	90.00%	–	–	–	–	–
Comotion-35 [50]	95.95%	85.35%	93.60%	88.25%	91.60%	–	–	–
Comotion-70 [50]	99.10%	93.25%	98.30%	90.45%	–	–	–	–
ADDNet-3d [65]	92.14%	83.93%	92.50%	78.21%	90.36%	78.21%	80.00%	69.29%
Long-Distance Attention [31]	**99.29%**	99.64%	98.58%	**94.29%**	–	–	–	–
MSVT [60]	95.79%	93.72%	92.93%	92.24%	–	–	–	–
FAMM [24]	–	–	–	–	90.00%	**91.00%**	92.75%	**87.50%**
ST-ILIF(ours)	98.57%	**99.64%**	**99.29%**	93.93%	**96.43%**	88.93%	**94.64%**	76.79%

5.3.4 Input sequence length evaluation. The aforementioned experiments were conducted on input sequences of five consecutive frames, and the results validated the effectiveness of the proposed method and module. In addition, to verify the effect of different input sequence lengths on model performance, another set of experiments was conducted for validation. The input lengths for the TDM module and sequence stream were set to 3, 5, 7, and

9, and the experimental results are presented in Table 6. Notably, a positive correlation is not observed between input sequence length and model performance. The experimental results indicate that the optimal performance of the model is achieved when the input consists of three consecutive frames. Increasing the input sequence length up to nine frames did not yield a significant improvement in performance. This observation is attributed to the fact that, as the sequence length increases, the inter-frame differences gradually become more pronounced, whereas the differences between real and fake videos become progressively subtle, thereby restricting the model's ability to extract discriminative features of high complexity.

Although the performance of the models with seven and nine input frames is higher than that with five input frames, the improvement is not significant. Moreover, increasing the number of input frames inevitably leads to increased model complexity and computational costs. Considering practical factors, we believe that this improvement is not worthwhile. Therefore, we opt for the best-performing sequence stream with three input frames for comparison with the model having five input frames. The results are shown in Table 7 show that the comprehensive performance of the model with five input frames is better than that of three input frames. The results of the corresponding generalization performance tests are presented in Table 8, indicating that the model with three input frames is much weaker than the model with five input frames in terms of generalization ability. The reason for this phenomenon may be that as the input sequence length decreases, the model is more likely to overfit specific tampering methods, resulting in lower generalization performance.

Based on the aforementioned analysis, it is evident that the performance of the ST-ILIF model is significantly influenced by the length of the input sequence. To comprehensively evaluate the detection performance of the model within the dataset as well as the cross-dataset generalization ability, we chose the five-frame input sequence length as the final configuration and conducted subsequent comparative experiments with other state-of-the-art methods.

Table 6. Evaluation of different input sequence lengths for sequence streams.The evaluation metric is AUC.

Architecture	Input sequence length	FF++ c23
sequence stream	Three frame	**96.57%**
	Five frame	95.71%
	Seven frame	96.00%
	Nine frame	95.86%

Table 7. Comparison of within-dataset detection performance between three-frame input ST-ILIF and five-frame input ST-ILIF.The evaluation metric is AUC and ACC. ACC represents the proportion of correctly classified samples to the total number of samples, which reflects the overall accuracy of the model in classification.

Models	FF++ c23 AUC	FF++ c23 ACC
ST-ILIF with three-frame input.	99.22%	97.14%
ST-ILIF with five-frame input.	**99.35%**	**97.29%**

5.4 Within-Dataset Evaluation

The performances of the various methods were compared on the FF++ dataset videos with different compression levels, using video-level accuracy as the evaluation metric. The obtained results are presented in Table 5.

Table 8. Comparison of generalization performance between three-frame input ST-ILIF and five-frame input ST-ILIF.The evaluation metric is AUC.

Models	Celeb-DF	DFDC	DF-1.0
ST-ILIF with three-frame input.	72.00%	**66.30%**	88.50%
ST-ILIF with five-frame input.	**78.95%**	63.37%	**91.07%**

The experimental results in Table 5 demonstrate that the proposed ST-ILIF achieves superior performance compared to most existing video detection methods. Although the porposed method is not as effective as the other methods for some subsets, it is lightweight and can achieve good results in limited environments. Although the number of parameters were not provided for the methods compared, reasonable inferences can be made from the backbone and algorithm used. Long-Distance Attention [31] uses EfficientNet [47] as the backbone and the original self-attention mechanism for feature aggregation. MSVT [60] uses three ResNet-50 [15] to extract different groups of subtle artifacts in parallel, and finally uses the original self-attention mechanism for fusion. The proposed method only uses two ResNet-18 [15] to extract features in parallel, and there is no large tensor multiplication in the algorithm. The parameters of each backbone are listed in Table 9. From the peraspective of using backbone and tensor algorithms, it is reasonable to speculate that the porposed method is more lightweight.

Our method exhibits poor performance on the NT subset within the FF++ dataset. We attribute this to the unique tampering method employed in the NT subset, resulting in highly subtle artifacts. Additionally, the compression operation may contribute to the loss or blurring of artifact information. Our method exploits short-term spatiotemporal information differences to capture artifacts. However, extracting several frames for difference calculations may not yield discernible pixel differences, or if they do, the differences may be very subtle. So, these subtle features gradually diminish during the network training process, rendering them difficult to capture.e.But our method is not specifically designed for the NT subset. As shown in Table 5, our approach achieves promising results on other subsets. In comparison to previous works that focus solely on either temporal or spatial information, our proposed method integrates both temporal and spatial features. By leveraging complementary information, we effectively capture tampering traces in both space and time, leading to better performance across other datasets.

Table 9. BackBone parameters and throughput used by each method.

BackBone	**Parameters**	**FLOPS**
Long-distance Attention [31]	43M	4.2B
MSVT [60]	25.6M	4.1B
ST-ILIF(ours)	11.7M	1.8B

5.5 Cross-Dataset Evaluation

To assess the generalizability of ST-ILIF, we trained the model on the FF++ C40 dataset and evaluated its performance on raw data from both the Celeb-DF and DFDC datasets. The results are listed in Table 10. In evaluating the model's performance on the DeeperForensics-1.0 dataset, the same configuration as in [59] was employed. Specifically, the model was trained on the FF++ C23 dataset and subsequently evaluated using the std/rand test set. In both instances, AUC was used as the evaluation metric, and the results are presented in Table 11.

Based on the results presented in Table 10, the proposed method outperforms all other methods in terms of generalization performance on the Celeb-DF dataset, achieving a 1.3% performance improvement the current state-of-the-art detection methods [12]. For example, [11, 12] exploited the spatiotemporal inconsistency information of fake videos and obtaineds better generalization performance. However, their performance is slightly worse because they do not perform feature fusion well. The proposed ST-ILIF can effectively perform feature interaction and fusion and thus has the best performance. In contrast, on the more challenging DFDC dataset, the performances of all detectors, including the proposed method, decreased substantially. We believe this is because the DFDC dataset contains various unknown tampering methods, resulting in more complex scenes in the forged videos. As a result, the phase-aware flow may have biases in recognizing the artifacts in DFDC, making it difficult for the model to learn discriminative features. Additionally, the method proposed in this study only focuses on short-term spatiotemporal disparity information and does not model long-term video information, leading to poorer performance.

Compared to the approach proposed by Delving et al, which utilizes time-series methods, our method considers the relationship between time and space. Moreover, we further analyze spatial features in the frequency domain, identifying the phase spectrum as the component most susceptible to forgery. Leveraging the reconstructed phase spectrum for feature extraction and integrating it with temporal features through dual-stream feature fusion, our approach enhances the model's ability to recognize tampering features, offering new insights for future work. Additionally, our method achieves superior results on the Celeb-DF dataset, which involves more post-processing operations and higher resolution.

To better match real-world scenes, DeeperForensics-1.0 adds various types of interference, such as color, saturation, and contrast transformations, to the std/rand subset. As shown in Table 11, the proposed method achieves a 6.86% performance improvement compared to [59]. This indicates that ST-ILIF has efficient spatiotemporal inconsistency modeling capability and is robust common image interference operations.

Table 10. The generalization evaluation on the Celeb-DF and DFDC datasets is presented, with results for some other methods taken from [12].The evaluation metric is AUC.

Methods	Celeb-DF	DFDC
Xception [44]	65.50%	59.39%
I3D [5]	74.11%	66.87%
VA-LogReg [34]	55.10%	–
TEI [30]	74.66%	67.42%
D-FWA [22]	56.90%	–
Capsule [36]	57.50%	–
V4D [62]	70.08%	67.34%
DIANet [17]	70.40%	–
DSANet [56]	73.71%	68.08%
ADDNet-3D [65]	60.85%	65.89%
STIL [11]	75.58%	67.88%
Delving [12]	77.65%	**68.43%**
ST-ILIF(ours)	**78.95%**	63.37%

Table 11. The performance of various methods in terms of generalization is evaluated on the "std/rand" subset of the DF-1.0 dataset. The comparison results with some other methods are obtained from [59].The evaluation metric is AUC.

Methods	FF++ c23	DF-1.0 std/rand
MesoNet [2]	98.76%	57.66%
Xception [44]	98.94%	70.03%
Unmasking [9]	85.44%	52.89%
TSDA [3]	96.00%	73.49%
Ensemble [4]	99.11%	73.45%
MTD-Net [59]	**99.38%**	84.21%
ST-ILIF(ours)	99.35%	**91.07%**

5.6 Visualization and Analysis

To acquire a more profound understanding of the complementary interplay between the two streams of ST-ILIF, we leveraged the gradient-weighted class activation mapping (Grad-CAM) approach [45] to visualize the feature maps. The visualization results are shown in Fig. 12.

For real video sequences, the phase-aware stream exhibits higher activation values around the mouth region, whereas the sequence stream pays more attention to the eyes and surrounding area. By contrast, in deepfake videos, the activation values of the phase-aware stream are generally low and tend to focus more on the edges of the face, whereas the sequence stream is more sensitive to changes in the central area of the face. These findings indicate that the two streams of ST-ILIF can provide complementary discriminative features for identifying deepfake videos, and the model can extract tampering traces from the spatiotemporal domain. In addition, the visualization results of the phase-aware stream show that the model focuses more on structural features after removing the image amplitude component. The smooth activation values of the sequence stream in the temporal domain demonstrate that the TDM module can learn spatiotemporally inconsistent features of real and fake videos.

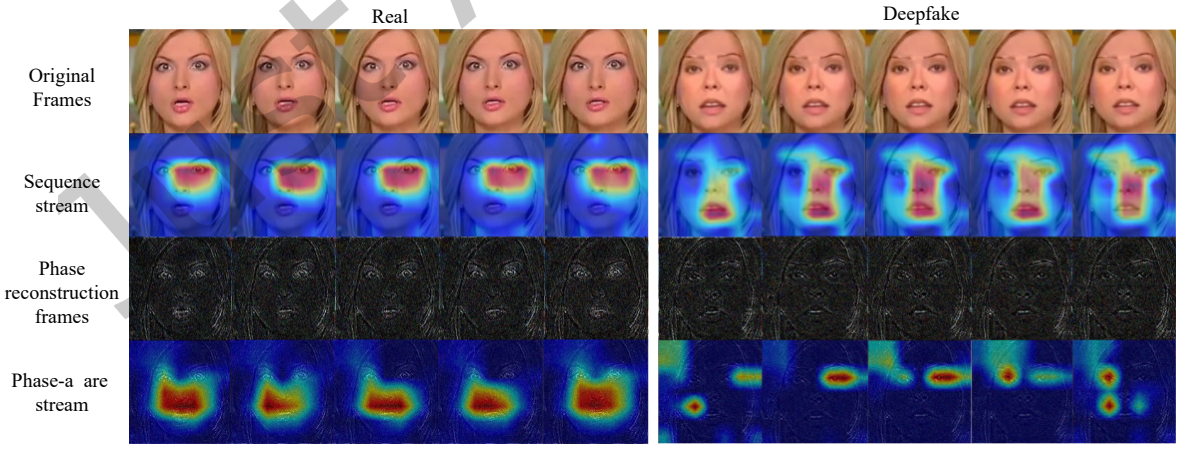

Fig. 12. Results of the feature map visualization of the phase-aware and sequence streams of ST-ILIF.

6 Conclusions and Future Work

Facing the abuse of deepfake technology, the declining quality of information within the metaverse. This paper presented a novel deepfake detection method called, ST-ILIF, which operates at the video level. The method leverages the spatiotemporal inconsistency cues left by deepfake generation techniques, which are detected by the phase-aware and sequence streams. To highlight the spatial inconsistency of deepfake videos, we proposed a phase-based frame reconstruction method that enhances the structural features of images, designing a phase-aware stream to capture spatially inconsistentcy features. The sequence stream utilized the TDM module to capture spatiotemporal inconsistency information from the input video sequence. Finally, the CAFIM and AFFM modules interactively fused the features of the two streams to adaptively amplify the tampering traces. A comprehensive evaluation was conducted on mainstream public datasets, demonstrating that the proposed ST-ILIF surpasses most existing methods in both in-dataset detection and cross-dataset generalization performance. Currently, the major challenge in the field of deepfake detection is insufficient generalization performance. To develop more universal deepfake detectors, our future work will focus on designing models that can extract long-term temporal features, adopting transformers for the design of the detection model.

7 Acknowledgments

This work was funded in part by the National Natural Science Foundation of China under Grant 62172059, 62272160, U22A2030 and 61972142, in part by Scientific Research Fund of Hunan Provincial Education Department of China under Grant 22A0200, and in part by Hunan Provincial Funds for Distinguished Young Scholars under Grant 2024JJ2025.

References

[1] 2019. Deepfakes github. https://github.com/deepfakes/faceswap

[2] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2018. MesoNet: a Compact Facial Video Forgery Detection Network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. 1–7. https://doi.org/10.1109/WIFS.2018.8630761

[3] Luca Bondi, Edoardo Daniele Cannas, Paolo Bestagini, and Stefano Tubaro. 2020. Training Strategies and Data Augmentations in CNN-based DeepFake Video Detection. In *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*. 1–6. https://doi.org/10.1109/WIFS49906.2020.9360901

[4] Nicolò Bonettini, Edoardo Daniele Cannas, Sara Mandelli, Luca Bondi, Paolo Bestagini, and Stefano Tubaro. 2021. Video Face Manipulation Detection Through Ensemble of CNNs. In *2020 25th International Conference on Pattern Recognition (ICPR)*. 5012–5019. https://doi.org/10.1109/ICPR48806.2021.9412711

[5] Joao Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6299–6308.

[6] M. del Castillo. 2022, September 1. Facebook's Metaverse Could Be Overrun By Deep Fakes And Other Misinformation If These Non-Profits Don't Succeed. https://www.forbes.com/sites/michaeldelcastillo/2022/08/29/facebooks-metaverse-could-be-overrun-by-deep-fakes-and-other-misinformation-if-these-non-profits-dont-succeed/?sh=21acb3842737.

[7] Xiangling Ding, Wenjie Zhu, and Dengyong Zhang. 2022. DeepFake Videos Detection via Spatiotemporal Inconsistency Learning and Interactive Fusion. In *2022 19th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 425–433.

[8] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. 2019. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854* (2019). https://doi.org/10.48550/arXiv.1910.08854

[9] Ricard Durall, Margret Keuper, Franz-Josef Pfreundt, and Janis Keuper. 2019. Unmasking deepfakes with simple features. *arXiv preprint arXiv:1911.00686* (2019). https://doi.org/10.48550/arXiv.1911.00686

[10] Shiming Ge, Fanzhao Lin, Chenyu Li, Daichi Zhang, Weiping Wang, and Dan Zeng. 2022. Deepfake video detection via predictive representation learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 18, 2s (2022), 1–21.

[11] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. 2021. Spatiotemporal Inconsistency Learning for DeepFake Video Detection. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3473–3481.

[12] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, and Lizhuang Ma. 2022. Delving into the Local: Dynamic Inconsistency Learning for DeepFake Video Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 36. 744–752.

[13] Zhiqing Guo, Gaobo Yang, Jiyou Chen, and Xingming Sun. 2021. Fake face detection via adaptive manipulation traces extraction network. *Computer Vision and Image Understanding* 204 (2021). https://doi.org/10.1016/j.cviu.2021.103170

[14] Bruce C Hansen and Robert F Hess. 2007. Structural sparseness and spatial phase alignment in natural scenes. *JOSA A* 24, 7 (2007), 1873–1885.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.

[16] Juan Hu, Xin Liao, Wei Wang, and Zheng Qin. 2022. Detecting Compressed Deepfake Videos in Social Networks Using Frame-Temporality Two-Stream Convolutional Network. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 3 (2022), 1089–1102. https://doi.org/10.1109/TCSVT.2021.3074259

[17] Ziheng Hu, Hongtao Xie, Yuxin Wang, Jiahong Li, Zhongyuan Wang, and Yongdong Zhang. 2021. Dynamic inconsistency-aware deepfake video detection. In *IJCAI*. 736–742.

[18] Gengyun Jia, Meisong Zheng, Chuanrui Hu, Xin Ma, Yuting Xu, Luoqi Liu, Yafeng Deng, and Ran He. 2021. Inconsistency-Aware Wavelet Dual-Branch Network for Face Forgery Detection. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3, 3 (2021), 308–319. https://doi.org/10.1109/TBIOM.2021.3086109

[19] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. 2020. 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 (CVPR)*. 2889–2898.

[20] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014). https://doi.org/10.48550/arXiv.1412.6980

[21] Dingquan Li, Tingting Jiang, and Ming Jiang. 2019. Quality Assessment of In-the-Wild Videos. In *Proceedings of the 27th ACM International Conference on Multimedia* (Nice, France) *(MM '19)*. Association for Computing Machinery, New York, NY, USA, 2351–2359. https://doi.org/10.1145/3343031.3351028

[22] Yuezun Li and Siwei Lyu. 2018. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656* (2018). https://doi.org/10.48550/arXiv.1811.00656

[23] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2020. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3207–3216.

[24] Xin Liao, Yumei Wang, Tianyi Wang, Juan Hu, and Xiaoshuai Wu. 2023. FAMM: Facial Muscle Motions for Detecting Compressed Deepfake Videos over Social Networks. *IEEE Transactions on Circuits and Systems for Video Technology* (2023).

[25] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. 2022. Robust high-resolution video matting with temporal guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 238–247.

[26] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. 2021. Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 772–781.

[27] Jiarui Liu, Kaiman Zhu, Wei Lu, Xiangyang Luo, and Xianfeng Zhao. 2021. A lightweight 3D convolutional neural network for deepfake detection. *International Journal of Intelligent Systems* 36, 9 (2021), 4990–5004. https://doi.org/10.1002/int.22499 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/int.22499

[28] Kunlin Liu, Wenbo Zhou, Zhenyu Zhang, Yanhao Ge, Hao Tang, Weiming Zhang, and Nenghai Yu. 2023. Measuring the Consistency and Diversity of 3D Face Generation. *IEEE Journal of Selected Topics in Signal Processing* 17, 6 (2023), 1208–1220. https://doi.org/10.1109/JSTSP.2023.3273781

[29] Xiaolong Liu, Yang Yu, Xiaolong Li, Yao Zhao, and Guodong Guo. 2023. TCSD: Triple complementary streams detector for comprehensive deepfake detection. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 19, 6 (2023), 1–22.

[30] Zhaoyang Liu, Donghao Luo, Yabiao Wang, Limin Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Tong Lu. 2020. Teinet: Towards an efficient architecture for video recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11669–11676.

[31] Wei Lu, Lingyi Liu, Bolin Zhang, Junwei Luo, Xianfeng Zhao, Yicong Zhou, and Jiwu Huang. 2023. Detection of Deepfake Videos Using Long-Distance Attention. *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[32] Fuyan Ma, Bin Sun, and Shutao Li. 2021. Facial Expression Recognition with Visual Transformers and Attentional Selective Fusion. *IEEE Transactions on Affective Computing* (2021). https://doi.org/10.1109/TAFFC.2021.3122146

[33] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. 2020. Two-Branch Recurrent Network for Isolating Deepfakes in Videos. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 667–684.

[34] Falko Matern, Christian Riess, and Marc Stamminger. 2019. Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*. 83–92. https://doi.org/10.1109/WACVW.2019.00020

[35] Changtao Miao, Qi Chu, Weihai Li, Suichan Li, Zhentao Tan, Wanyi Zhuang, and Nenghai Yu. 2022. Learning Forgery Region-Aware and ID-Independent Features for Face Manipulation Detection. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 4, 1 (2022), 71–84. https://doi.org/10.1109/TBIOM.2021.3119403

[36] Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. 2019. Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2307–2311. https://doi.org/10.1109/ICASSP.2019.8682602

[37] Yuval Nirkin, Yosi Keller, and Tal Hassner. 2019. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*. 7184–7193.

[38] A Oppenheim, Jae Lim, Gary Kopec, and SC Pohlig. 1979. Phase in speech and pictures. In *ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 4. IEEE, 632–637.

[39] Alan V Oppenheim and Jae S Lim. 1981. The importance of phase in signals. *Proc. IEEE* 69, 5 (1981), 529–541.

[40] Guilin Pang, Baopeng Zhang, Zhu Teng, Zige Qi, and Jianping Fan. 2023. MRE-Net: Multi-Rate Excitation Network for Deepfake Video Detection. *IEEE Transactions on Circuits and Systems for Video Technology* (2023). https://doi.org/10.1109/TCSVT.2023.3239607

[41] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, et al. 2020. DeepFaceLab: Integrated, flexible and extensible face-swapping framework. *arXiv preprint arXiv:2005.05535* (2020). https://doi.org/10.48550/arXiv.2005.05535

[42] Leon N Piotrowski and Fergus W Campbell. 1982. A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. *Perception* 11, 3 (1982), 337–346.

[43] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. 2020. Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 86–103.

[44] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 1–11.

[45] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 618–626.

[46] Saniat Javid Sohrawardi, Akash Chintha, Bao Thai, Sovantharith Seng, Andrea Hickerson, Raymond Ptucha, and Matthew Wright. 2019. Poster: Towards Robust Open-World Detection of Deepfakes. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security* (London, United Kingdom) *(CCS '19)*. Association for Computing Machinery, New York, NY, USA, 2613–2615. https://doi.org/10.1145/3319535.3363269

[47] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR, 6105–6114.

[48] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2387–2395.

[49] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features With 3D Convolutional Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 4489–4497.

[50] Gengxing Wang, Jiahuan Zhou, and Ying Wu. 2020. Exposing Deep-faked Videos by Anomalous Co-motion Pattern Detection. *arXiv preprint arXiv:2008.04848* (2020). https://doi.org/10.48550/arXiv.2008.04848

[51] Hanyi Wang, Zihan Liu, and Shilin Wang. 2023. Exploiting Complementary Dynamic Incoherence for DeepFake Video Detection. *IEEE Transactions on Circuits and Systems for Video Technology* (2023). https://doi.org/10.1109/TCSVT.2023.3238517

[52] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. 2021. TDN: Temporal Difference Networks for Efficient Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1895–1904.

[53] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. 2020. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 11531–11539. https://doi.org/10.1109/CVPR42600.2020.01155

[54] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. 2020. CNN-Generated Images Are Surprisingly Easy to Spot... for Now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8695–8704.

[55] Tianyi Wang, Harry Cheng, Kam Pui Chow, and Liqiang Nie. 2023. Deep convolutional pooling transformer for deepfake detection. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 19, 6 (2023), 1–20.

[56] Wenhao Wu, Yuxiang Zhao, Yanwu Xu, Xiao Tan, Dongliang He, Zhikang Zou, Jin Ye, Yingying Li, Mingde Yao, Zichao Dong, et al. 2021. Dsanet: Dynamic segment aggregation network for video-level representation learning. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1903–1911.

[57] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. 2021. A Fourier-Based Framework for Domain Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14383–14392.

[58] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. 2021. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14383–14392.

[59] Jiachen Yang, Aiyun Li, Shuai Xiao, Wen Lu, and Xinbo Gao. 2021. MTD-Net: Learning to Detect Deepfakes Images by Multi-Scale Texture Difference. *IEEE Transactions on Information Forensics and Security* 16 (2021), 4234–4245. https://doi.org/10.1109/TIFS.2021.3102487

[60] Yang Yu, Rongrong Ni, Yao Zhao, Siyuan Yang, Fen Xia, Ning Jiang, and Guoqing Zhao. 2023. MSVT: Multiple Spatiotemporal Views Transformer for DeepFake Video Detection. *IEEE Transactions on Circuits and Systems for Video Technology* (2023).
[61] Dengyong Zhang, Jiahao Chen, Xin Liao, Feng Li, Jiaxin Chen, and Gaobo Yang. 2024. Face Forgery Detection via Multi-Feature Fusion and Local Enhancement. *IEEE Transactions on Circuits and Systems for Video Technology* (2024), 1–1. https://doi.org/10.1109/TCSVT.2024.3390945
[62] Shiwen Zhang, Sheng Guo, Weilin Huang, Matthew R Scott, and Limin Wang. 2020. V4d: 4d convolutional neural networks for video-level representation learning. *arXiv preprint arXiv:2002.07442* (2020). https://doi.org/10.48550/arXiv.2002.07442
[63] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. 2021. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2185–2194.
[64] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Weiming Zhang, and Nenghai Yu. 2022. Self-supervised transformer for deepfake detection. *arXiv preprint arXiv:2203.01265* (2022).
[65] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. 2020. *WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection*. Association for Computing Machinery, New York, NY, USA, 2382–2390.

Received 29 December 2023; revised 6 April 2024; accepted 4 May 2024