Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.062922

ARTICLE

Deepfake Detection Method Based on Spatio-Temporal Information Fusion

Xinyi Wang*, Wanru Song, Chuanyan Hao and Feng Liu

Department of Digital Media Technology, Nanjing University of Posts and Telecommunications, Nanjing, 210023, China

*Corresponding Author: Xinyi Wang. Email: xawangxy@njupt.edu.cn

Received: 31 December 2024; Accepted: 25 February 2025; Published: 16 April 2025

ABSTRACT: As Deepfake technology continues to evolve, the distinction between real and fake content becomes increasingly blurred. Most existing Deepfake video detection methods rely on single-frame facial image features, which limits their ability to capture temporal differences between frames. Current methods also exhibit limited generalization capabilities, struggling to detect content generated by unknown forgery algorithms. Moreover, the diversity and complexity of forgery techniques introduced by Artificial Intelligence Generated Content (AIGC) present significant challenges for traditional detection frameworks, which must balance high detection accuracy with robust performance. To address these challenges, we propose a novel Deepfake detection framework that combines a two-stream convolutional network with a Vision Transformer (ViT) module to enhance spatio-temporal feature representation. The ViT model extracts spatial features from the forged video, while the 3D convolutional network captures temporal features. The 3D convolution enables cross-frame feature extraction, allowing the model to detect subtle facial changes between frames. The confidence scores from both the ViT and 3D convolution submodels are fused at the decision layer, enabling the model to effectively handle unknown forgery techniques. Focusing on Deepfake videos and GAN-generated images, the proposed approach is evaluated on two widely used public face forgery datasets. Compared to existing state-of-the-art methods, it achieves higher detection accuracy and better generalization performance, offering a robust solution for deepfake detection in real-world scenarios.

KEYWORDS: Deepfake detection; vision transformer; spatio-temporal information

1 Introduction

The rapid development of deep learning technology has significantly improved the efficiency of audio and video processing, with artificial intelligence (AI) algorithms playing a key role in this process. AI has become an indispensable part of various industries, however, it has two sides. While it brings great convenience, the risks caused by its misuse should not be overlooked. In recent years, the use of AI technology for illegal activities has emerged as a growing concern. Among these, Deepfake technology [1] has attracted significant attention.

Deepfake, a combination of "deep learning" and "fake", utilizes Generative Adversarial Networks (GANs) [2] to generate fake content, such as face swapping [3–5] and other effects. For example, in the film and television industry, such as the post-production of Avatar, which utilizes this technology to present better visual and sensory effects for the audience. Although this technology has shown great potential in media, education and other fields, its malicious use has also caused serious social problems. A report released in October 2019 by Dutch cybersecurity startup Deeptrace estimated that 96 percent of all online Deepfakes are pornographic. In addition to creating fake pornographic videos, the technique has also been used to

misrepresent well-known politicians in videos. Consequently, efficient detection methods for Deepfake videos have become a key research topic [6].

2 Structure

Early forgery detection methods could identify forged faces to some extent. For example, Popescu et al. [7] proposed detecting double JPEG compression in images by analyzing whether the first-order statistics of DCT coefficients exhibit specific artifacts. Similarly, Matern et al. [8] introduced a forgery detection method based on facial feature extraction, focusing on characteristics such as pupil color, eye reflections, and blurriness in tooth details. However, these methods rely heavily on noticeable image defects and feature extraction, which may limit their effectiveness when dealing with more advanced forgery techniques and complex application scenarios.

The existing deep learning methods for detecting Deepfake videos are mostly based on spatial features at the frame level of the video. These methods detect inconsistencies between the source and fake videos, either physiologically or physically, or identify artifacts generated during video synthesis, as well as easily detectable traces of forgery. Depending on the focus of attention, Deepfake forgery detection techniques can currently be divided into three types.

1. Detecting inconsistencies in the physiological and physical aspects of Deepfake videos. For example, Pishori et al. [9] proposed detecting eye blinking in the face of a fake video to determine whether the video is fake. Guarnera et al. [10] used fingerprints left during the image generation process for identification. Ciftci et al. [11] suggested detecting hidden authentic biological signals in fake videos to identify differences in time and spatial features exhibited by the forgery. Li et al. [12] observed abnormal blinking behavior in synthesized videos.
2. Detecting the authenticity of DeepFake videos by utilizing signal-level artifacts generated during the synthesis process. Li et al. [13] proposed a FaceX-ray method for detecting differences in image boundaries. Li et al. [14] proposed a semi-supervised method to identify forged videos, but this method cannot be applied to the entire video and can only detect forgeries frame-by-frame. Zhu et al. [15] combined direct facial illumination and identity texture as cues for detecting subtle traces of forgeries. MesoNet [16] is capable of automatically detecting facial forgery in videos. By thoroughly analyzing and constructing a lightweight network that focuses on mesoscopic features of images, it achieves efficient identification of forged faces. However, its limitation lies in being applicable only to specific forgery techniques.
3. Combing the source videos to obtain the traces of forgery that appear in the DeepFake videos. Chugh et al. [17] proposed an approach based on audio-video combination to detect the difference between real and fake videos. However, the applicability of this approach is limited in scenarios where not all datasets have audio sources; Qin et al. [18] demonstrated that the DeepFake generation process introduces AI-specific traces, particularly around the mouth and eyes. Nguyen et al. [19] introduced the capsule network method, which can detect post-recording attacks on generated videos. Most of these methods rely solely on information within individual frames, performing frame-by-frame analysis without leveraging the temporal information in videos. The multi-task learning network [20] was primarily effective against facial reenactment attacks and face swapping attacks. To enhance the network's generalization capability, a semi-supervised learning approach was employed. Nirkin et al. [21] proposed a method for DeepFake detection based on discrepancies between faces and their context, but it also has certain limitations. Zhao et al. [22] proposed a multi-attention-based DeepFake detection network, treating DeepFake detection as a fine-grained classification problem. By combining texture enhancement blocks and bilinear attention pooling, the network effectively extracts features.

M2TR [23] combined a multi-modal multi-scale Transformer structure with frequency domain feature detection, enabling the detection of forgery traces at different scales in images. In 2024, Keresh et al. [24] tackled the challenge of Deepfake video detection by leveraging the DINO framework to fine-tune a Vision Transformer (ViT) model. This method allows the model to identify distinguishing features in unlabeled data, improving its ability to detect manipulated content.

In summary, existing Deepfake detection methods have made significant optimizations to address the characteristics of forged videos. However, these methods struggle to fully explore the forgery traces in both temporal and spatial dimensions. Additionally, the modeling of spatiotemporal inconsistencies lacks global relationship representation, resulting in issues of insufficient accuracy and poor generalization. Therefore, effectively utilizing the spatiotemporal inconsistency information in forged videos to obtain complete and high-quality feature representations is crucial for Deepfake video detection. To solve these problems, this paper proposes a novel Deepfake detection method that leverages the spatiotemporal information from 3D ConvNet [25] and Vision Transformer (ViT) [26] to enhance feature representation and improve the model's generalization ability. This method optimizes the extraction process of dynamic and temporal features from video data with the following three main contributions.

1. Considering the entire video sequences as a whole, 3D ConvNet is utilized to integrate the spatial information and temporal features of the video to fully explore the spatio-temporal forgery traces.
2. Taking advantage of ViT in long-range modeling, it is combined with 3D ConvNet to capture global spatio-temporal inconsistencies and further enhance feature characterization.
3. The method mines more generalized forensic patterns in Deepfake video detection, which improves the accuracy and generalizability of the detection.

3 Related Studies

3.1 Deepfake Generation Methods

The first public Deepfake product is FakeApp [27], created by a Reddit user based on the Autoencoder-Decoder structure [28]. This method uses an autoencoder to extract features from compressed facial images and a decoder to reconstruct the images. The decoder can generate clear facial images from noisy inputs. To swap faces between source and target images, two encoder-decoder pairs are used. Each pair is trained on a separate facial dataset, and the encoder parameters are shared between the two pairs. Similar methods are also used in DeepFaceLab [29] and DFaker [30]. Recent advancements in deepfake detection include methods that utilize multimodal data. For instance, Salvi et al. [31] proposed a robust multimodal approach to deepfake detection, integrating both visual and non-visual cues. Additionally, Sunanda et al. [32] introduced the use of CNNs to identify manipulated visual media, demonstrating significant improvements in detection accuracy across various datasets.

The autoencoder architecture is an unsupervised learning neural network model designed to learn a compressed representation of the input data and then reconstruct the original input data from this encoding. Two convolutional neural networks (CNNs) are used as encoder and decoder respectively. The encoder part encodes the input data and projects the information into a low-dimensional potential space, which summarizes the key features of the image. The decoder part reconstructs the image from the information in the low-dimensional potential space. By making appropriate modifications to the latent representation, editing of the image, such as changing features like expression, age, gender, etc., can be realized. As shown in Fig. 1, the decoders of the original and target images share the same encoder to generate different feature labels. The shared encoder completes the reconstruction process of face data by pairing different decoder combinations.

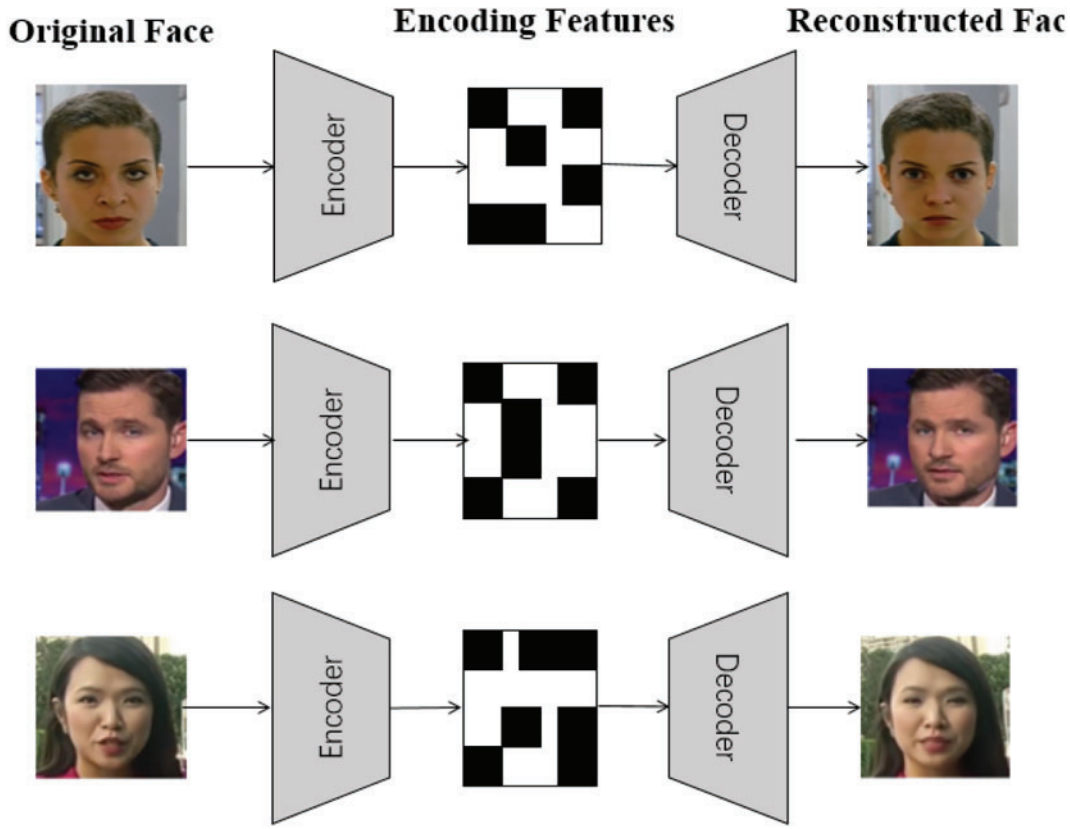

Figure 1: Face manipulation using Autoencoder-Decoder architecture

3.2 Vision Transformer

Since the introduction of the Transformer [33], it has long dominated the field of natural language processing (NLP), and its success has sparked widespread interest in exploring its application in other areas. For a long time, scientists have sought the unity of things in objective laws, attempting to explain the various complex problems in the natural world using more unified principles. This concept of unity is also reflected in the field of artificial intelligence. The field of computer vision has traditionally relied on convolutional neural networks (CNNs) to process image data. However, with the growing awareness of the need for model scalability and global relationship modeling, researchers began to consider whether the powerful capabilities of Transformers could be applied to the field of computer vision. This line of thinking led to a major innovation: Vision Transformer (ViT) [26]. Its introduction brought about a revolutionary transformation in the field of computer vision. Since 2020, Transformers have also demonstrated their strong capabilities in computer vision, indicating that the fields of natural language processing and computer vision may eventually converge.

ViT is an evolved version of the Transformer model successfully applied to image processing tasks, with several key differences from the original Transformer. Traditional Transformer models typically process input sequences consisting of text embeddings, whereas ViT adopts a unique approach to handle image information. Specifically, ViT divides the input image into fixed-size patches and rearranges them into a one-dimensional sequence. Each patch is treated as a "token" and is combined with a position vector representing its location in the image, along with a class vector for classification tasks. The core of the model is composed of the Transformer encoder, retaining the main multi-head attention structure but adjusting the position of

the normalization layers. In the final layer of the stacked Transformer encoders in ViT, the class vector in the sequence is considered to contain the global information of the image, thus serving as the representation of the entire image. This vector is passed through a fully connected layer to map the image representation to the output space for specific tasks.

In summary, the Vision Transformer (ViT) process begins with labeled facial images, which are segmented into patches representing local features. These patches are then input into the Transformer's encoding area, where position encoding and feature learning take place. After this, the position and feature information are classified, and the final result is obtained.

The ViT model consists of three parts: the Linear Projection of Flattened Patches (Embedding layer), which transforms the image data into a suitable form for processing; the Transformer Encoder, which learns the input image features; and the MLP Head, used for the final classification task.

For an image $x \in R^{H\times W\times C}$, it is first divided into N patches of size $M \times M$, where H, W, C represent the height, width, and number of channels of the image, respectively. M is the width or height of each patch, and N = HW/M^2. These N patches are then flattened along the channel dimension to form a 2D sequence $x_p \in R^{N\times M^2C}$. The 2D sequence x_p undergoes patch embedding, where each vector is linearly transformed, reducing the flattened 2D sequence into a D-dimensional vector $x_0 = x_p\mathbb{E}, \mathbf{E} \in R^{M^2C\times D}$. Furthermore, Vision Transformer embeds a learnable class token x_{cdass} at the beginning of the 2D sequence to eliminate the need to choose which flattened 2D embedding vector is used for the final classification prediction. A learnable positional embedding $\mathbb{E}_{p\varepsilon}$ is also added to include position information in the sequence. The final input sequence is as shown in Formula (1):

$$z_0 = \left[x_{class}; x_p^1\mathbf{E}; x_p^1\mathbf{E}; \ldots; x_p^N\mathbf{E}\right] + \mathbf{E}_{pos} \tag{1}$$

The second part consists of L stacked Transformer Encoders. After receiving the input sequence, the model performs L iterations of feature learning to obtain both global and local image features. Finally, the classification prediction result is obtained by applying MLP and normalization processing to the x_{class} of z_0.

Vision Transformer has demonstrated strong capability in detecting deepfake videos by learning the subtle changes in facial details within specific scenes. It effectively identifies visual artifacts in deepfake videos, making it well-suited for detecting forgery traces. For deepfake detection, particularly regarding facial details, this paper utilizes Vision Transformer to capture and analyze these features.

3.3 D ConvNet

Convolutional operations typically handle two-dimensional features of an image; however, CNNs are limited in capturing the spatial features of depth-forged videos, as they fail to account for the temporal information. To address this, temporal features must be incorporated into the 2D features of the video, transforming the CNN into a 3D CNN. This enables the model to capture both spatial and temporal transformations within the video. By stacking consecutive frames of depth-forged videos, 3D convolution is performed across these frames, allowing the model to recognize object motion over time. As shown in Fig. 2, the feature map is derived by convolving three consecutive frames at the same position, focusing on their local receptive fields.

While 2D convolution is commonly used for analyzing individual frames, it often neglects temporal information. As illustrated in Fig. 2a, 2D convolution learns the features of a single frame. In contrast, 3D CNNs capture both spatial features of each frame and temporal relationships between consecutive frames, as shown in Fig. 2b. This highlights how a 3D CNN learns spatiotemporal features in depth-forged video sequences.

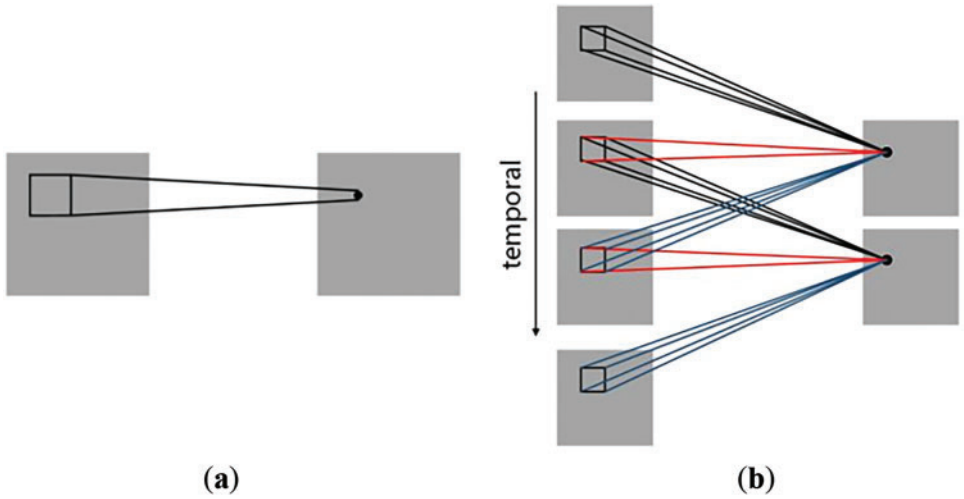

Figure 2: The process of feature learning: (a) 2DCNN performs image convolution using a 2D convolution kernel, (b) 3D CNN uses a 3D convolution kernel for convolution of image sequences

3D convolution can only extract specific features from the cube formed by consecutive frames in depth-forged videos. This limitation arises because all convolution operations within the cube use the same fixed weights. As a result, the features learned by the convolution kernels are uniform, leading to a lack of diversity in feature extraction. To address this, multiple weight settings can be introduced to capture different features across various dimensions of the cube.

In comparison, 2D convolution lacks an additional channel for feature extraction, which results in the loss of some three-dimensional information in depth-forged videos. It can process either a sequence of consecutive video frames or different 2D slices of a 3D cube. On the other hand, 3D CNNs are well-suited for tasks like video classification and image segmentation. In the medical field, most data is inherently 3D. For example, while a slice viewed with the naked eye appears 2D, it is part of a 3D structure. Identifying pathological tissues, such as tumors, involves analyzing a 3D object through its 2D sections, where 3D CNNs can be effectively applied. For classification tasks, 3D CNNs capture both spatial and temporal features across consecutive frames, delivering superior performance. This makes them highly effective for applications in depth-forged video detection and medical imaging analysis.

4 Methods

With the continuous evolution of Deepfake, the distinction between real and fake data has become increasingly blurred. Currently, some mainstream Deepfake video detection methods suffer from low detection accuracy and weak generalization performance. To address these issues, this paper proposes a deep forgery video detection method based on spatio-temporal feature fusion. The model integrates a 3D CNN network and Vision Transformer (ViT) to effectively distinguish between real and fake face videos generated by various forgery algorithms.

4.1 Network Architecture

The model structure used in this paper is shown in Fig. 3, and its process is as follows. First, frames are extracted from the video, and facial regions are cropped to generate both video and optical flow sequences. The main framework of the model adopts a dual-branch network, where each branch specializes in capturing spatial and temporal features of forged videos while suppressing irrelevant ones. The model

aims to capture spatial forgery traces within video frames and temporal inconsistencies between frames, thus identifying representative spatiotemporal forgery artifacts for more accurate predictions. To achieve this, both feature types are aggregated into a strongly correlated spatiotemporal inconsistency feature sequence at each feature extraction stage. These feature sequences are then input into the Vision Transformer (ViT) module for feature interaction, long-distance dependency modeling, and obtaining global spatiotemporal inconsistency features.

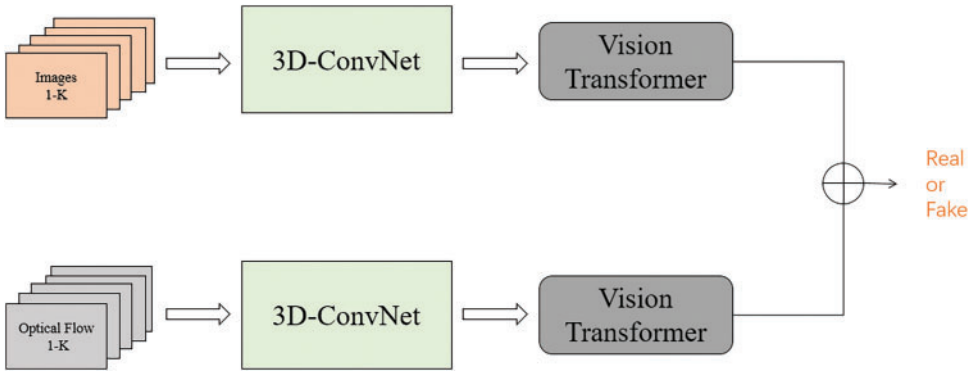

Figure 3: Model structure of the proposed method

4.2 Two-Stream Inflated 3D-ConvNet (I3D)

The 3DCNN network used in this paper is the Two-Stream Inflated 3D Convolutional Network (I3D) [34], which effectively learns the spatial and temporal features of deepfake facial videos. Originally proposed for action classification, the I3D network can learn continuous spatiotemporal features from videos. It is based on state-of-the-art image classification architectures, but with expanded filters and kernels, which are then projected into 3D, creating a deeply natural spatiotemporal classifier. The I3D employs a dual-stream approach for analyzing deepfake videos, with one stream processing RGB data and the other analyzing optical flow (FLOW) data. The results from the RGB and FLOW streams are fused. The flow images in the dual-stream structure represent the instantaneous motion of objects on a 2D plane during movement, capturing the correlation of behavior features between consecutive frames and detecting differences between them to compute frame correspondence. For deepfake videos, the temporal features between adjacent frames are inconsistent, irregular, and lack continuity, whereas real videos exhibit continuity and correlation between adjacent frames, with highly similar temporal features. Therefore, real and fake videos show significant feature differences.

The I3D network model is primarily derived from the Inception-V1 network, as shown in Fig. 4, which illustrates the Inception module and the 3D-Inception network model used to extract temporal information from video data. This model combines 2D convolutional network filters and pooling layers, adding a temporal dimension to better handle video data. The Inception module in the model includes three convolutional kernels of different sizes and a max-pooling layer. The model is trained separately on RGB data and optical flow data, and their respective predictions are averaged and fused at the end. The final classification result is output through a softmax function.

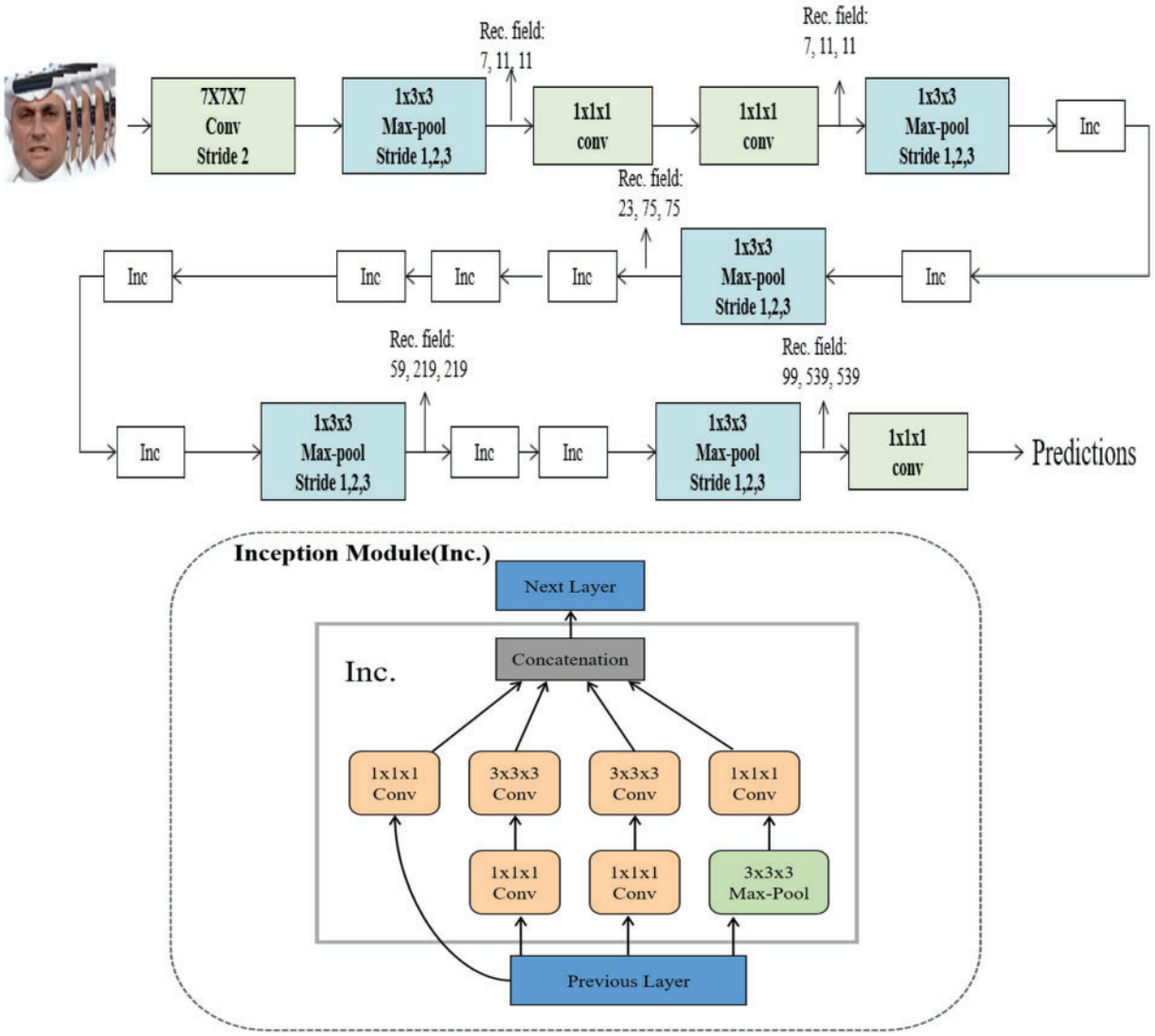

Figure 4: 3D-Inception

To validate the effectiveness of the optical flow information extracted by the model, a dual-stream architecture is designed, where both channels are trained using RGB data. The base structure of I3D is Inception-V1. I3D extends 2D convolutions into 3D, adding a temporal dimension to both convolutional and pooling layers. This is achieved by inflating all filters and pooling kernels from 2D to 3D, where the 2D filters' weights are repeated along the time dimension, normalized by dividing by N, and then rescaled. The 3D convolution enables the learning of temporal features, but the original flow branch still struggles to capture the fine-grained features of the face.

The main components of the Inflated Inception-V1 include convolutional layers, pooling layers, and Inc layers. In this network, the first convolutional layer has a stride of 2, followed by four max-pooling layers with a stride of 2. Before the final linear classification layer, there are four more max-pooling layers and one average-pooling layer. The model is trained using 300 frames of video at 25 frames per second and tested on the entire video.

4.3 The Function of the ViT Module in the Network

This study addresses the accurate extraction of forgery traces within individual video frames and their discontinuities across frames. By leveraging cross-domain information fusion, the interaction between spatial and temporal features is enhanced, improving the model's ability to analyze video data. As shown in Fig. 5, the Vision Transformer (ViT) model captures long-range dependencies by considering the relationships between different regions in an image. Under sufficient training, ViT outperforms state-of-the-art CNN models, suggesting a higher performance ceiling. However, ViT faces challenges such as limited locality, low sample efficiency, and difficulty in modeling complex visual features. In contrast, 3D CNNs excel at local feature extraction but may lose spatial information during convolution and pooling, reducing sensitivity to positional details. Thus, a hybrid model combining ViT and 3D CNN leverages the strengths of both, compensating for their weaknesses and providing more accurate representations.

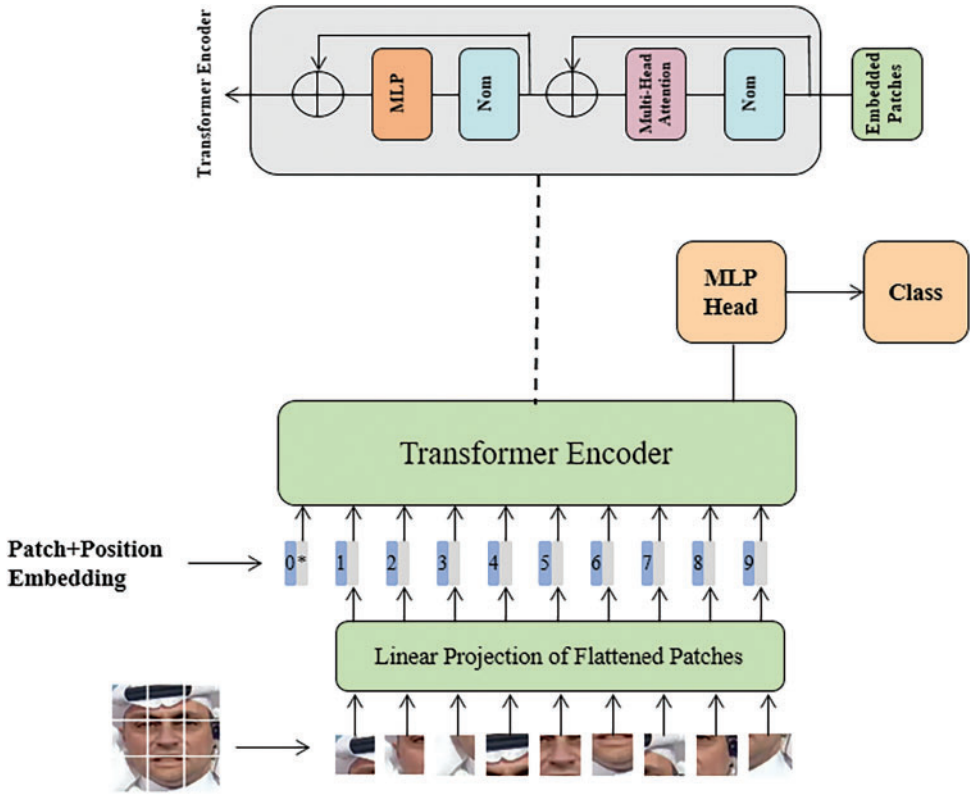

Figure 5: ViT module

To address the temporal feature neglect in existing methods for deepfake video detection, the proposed 3DCNN-ViT model integrates ViT for collaborative spatiotemporal feature learning. The ViT sub-model focuses on spatial features, effectively extracting semantic information from each frame. The 3D CNN sub-model captures temporal features by enabling cross-frame feature extraction, allowing it to detect subtle changes in the forged face. The confidence scores from both models are then fused at the decision layer for final classification.

Existing methods often overlook the spatial information's role in enhancing detection performance while extracting temporal features. This study proposes a hybrid model combining ViT and 3D CNN to retain spatial features while focusing on temporal information. By preprocessing video data with varying sampling

strides and combining spatial and temporal convolutions, we extract comprehensive spatiotemporal features from both high- and low-sampling-rate data. The ViT model then captures global relationships within the spatiotemporal feature sequence, further improving feature representation.

4.4 Necessity of Combining 3D ConvNet with ViT

While the Vision Transformer (ViT) excels at capturing spatial features through self-attention mechanisms, it primarily focuses on static spatial information and is not optimized for temporal features in videos. Since videos contain dynamic changes across consecutive frames, relying solely on ViT for temporal feature extraction presents certain limitations. To address this, we incorporate a 3D Convolutional Network (3D ConvNet). Unlike ViT, 3D ConvNet captures both spatial and temporal features by performing convolutions across multiple consecutive frames, enabling it to detect temporal inconsistencies between frames. This is essential for identifying dynamic forgeries in Deepfake videos. By combining the spatial feature learning of ViT with the spatiotemporal feature extraction capabilities of 3D ConvNet, our approach effectively leverages both spatial and temporal information, thereby improving the accuracy and robustness of deepfake detection in videos.

4.5 Loss Function

The loss function, also referred to as the cost function, maps the values of random events or variables related to random events to non-negative real numbers, representing the "risk" of these events. The fundamental task of deep learning is to approximate a function f(x) that maps input images to corresponding labels, with the loss function quantifying the quality of this mapping. Different loss functions exhibit varying performance outcomes. The method proposed in this paper is inherently a model for binary classification tasks. Hence, during the training process, cross-entropy loss is employed as the loss function. The mathematical form of the cross-entropy loss is as follows:

$$L_i = -y^T \log(P_i(Y_{p_i})) \tag{2}$$

$$P_i(Y_{p_i}) = \text{soft } max(cls_i(Y_{p_i})) \tag{3}$$

Among them, L_i is the loss of the i-th branch; y is the true label of the input sample, represented by a one-hot vector; cls_i represents the classifier of the i-th branch; P_i represents the classification probability calculated by the i-th branch. The final L_{total} is expressed as:

$$L_{\text{total}} = L_1 + L_2 \tag{4}$$

where L_1 is the loss of the video sequence branch, and L_2 represents the loss of the optical flow branch.

5 Experiments

The experiments in this study are conducted on an Ubuntu 16.04 operating system with the following hardware configuration: NVIDIA GeForce GTX 4070Ti, Intel Core i7-12700K CPU, 64 GB DDR4 RAM, and Python 3.7.

When extracting temporal features from video data, high sampling rate continuous video frames are essential, as low sampling rates may lead to insufficient extraction of temporal information. On the other hand, for spatial feature extraction, inputting high-sampling rate video frames can waste computational resources without necessarily improving detection performance. To more effectively extract spatiotemporal features, video data is preprocessed with different sampling rates based on specific needs. Low-sampling rate

video frames help enhance the model's ability to express spatial semantics, while high-sampling rate frames ensure the completeness of temporal features.

Specifically, video data is sampled with different time steps, t_S and t_T. The input video is represented as a continuous image sequence V, which is divided into V/t separate image sets. When t_S is set to 32, the model processes only one frame from every 32-frame segment. For different sampling intervals t, using a larger time step t_S ($t_S/8 = t_T$) during sampling allows for better utilization of the spatial features within the video frames. Similarly, reducing the sampling step and increasing the video sampling rate helps capture temporal inconsistencies between frames.

Moreover, Deepfake techniques typically focus on manipulating facial regions in videos. By cropping the background, the model's computational load can be reduced, while preventing overfitting to the background data. Therefore, RetinaFace is first used to preprocess the input data, locating the facial region based on anchor points and setting face bounding boxes. Additionally, each face region in the frames is enlarged by 1.5 times and cropped to a size of 320 × 320.

5.1 Dataset and Experimental Setup

In the real world, people's facial expressions and postures vary quite complexly, and these changes may be affected by multiple factors such as the environment, emotions, and lighting. At the same time, since the lighting, background, and environmental conditions in real scenes may change, the dataset used in the experiment should contain data samples of various facial expressions and postures and taken under different environmental conditions to ensure that the model can adapt to various environments. In addition, since detection in real scenes also needs to consider people of different races, genders, and ages, the dataset should cover a diverse population to ensure that the model can work effectively among a wide range of people. Using appropriate and rich datasets, deep fake detection models can be better generalized to various situations in the real world, improving their accuracy, generalization, and practicality. Therefore, the experiment selected two widely used, richly typed, large-scale, high-quality datasets. The information of these two datasets is shown in Table 1.

Table 1: Experimental datasets

Dataset	Quantity	Intra-dataset training	Cross-dataset evaluation (%)
FaceForensics++ [35]	5000	√	100
Celeb-DF [36]	5639	√	100

The FaceForensics++ (FF++) [35] dataset is a large-scale deep fake face video dataset used to study deep fake detection and face tampering detection technology. It is one of the most widely used datasets. This dataset is a large face dataset shared by Google. It collects 1000 initial videos from the YouTube video network, including a dataset of more than 1.8 million images. These videos have been forged using four face tampering methods, namely Deepfakes, Face2Face, FaceSwap, and NeuralTextures. Among them, DeepFake is based on the well-known GAN technology. For the faces it generates, the recognition rate of the human eye is about 75%. The FaceSwap method uses a deep learning method to completely reconstruct the face. It can use the model to swap the target face image. The recognition rate of the human eye for this algorithm is also around 75%. The Face2Face method replaces the target face with the real face by swapping it. There is no new face. In this case, the recognition rate of the human eye is only 41%. Each video contains a frontal face and is traceable. This is a well-known large dataset that can be used in face forensics.

The Celeb-DF [36] dataset takes into account people of different genders, ages, and races. It collects 590 real videos of 59 celebrities from YouTube, and then uses the DeepFakes method to generate 5639 fake videos in MPEG4.0 format with an average length of 13 s. Since this dataset can improve the face resolution, establish a color conversion algorithm for the faces in the fake video and the original video, and better integrate the boundaries of the fake area and the original area, it greatly improves the quality of the generated fake data and can be used to simulate fake generated videos in real environments.

Since the forged areas of the samples are mostly concentrated in the facial features, in order to make the model pay more attention to the characteristics of the forged areas, in the data preprocessing stage, 32 frames were captured at equal time intervals for each video in the FaceForensics++ and Celeb-DF datasets and processed as key frames as experimental samples. Then, the MTCNN face detection algorithm was used to locate the face in each frame to determine the facial rectangle. Finally, after face alignment, the image was cropped to a size of 320 dpi × 320 dpi, and divided into training set, validation set and test set in a ratio of 6:2:2.

For the training parameters in the network, the learning rate is 0.001. After about 200 iterations, the loss begins to converge. The activation function is the Relu function, the loss function is the cross entropy loss function, and the epoch is set to 200. The model uses the AdamW optimization algorithm to dynamically adjust the learning rate. The initial learning rate is set to 1×10^{-4}, the training batch size is set to 18, and the epoch is set to 40.

5.2 Evaluation

Video face forgery detection is essentially a binary classification problem. In this paper, the following evaluation metrics are used to comprehensively evaluate the performance of the proposed method:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \tag{7}$$

TP is the number of samples correctly predicted as positive, which means true positive examples. TN is the number of samples correctly predicted as negative, which means true negative examples. FN is the number of samples incorrectly predicted as negative, which means false negative examples. FP is the number of samples incorrectly predicted as positive, which means false positive examples.

In addition to these indicators, the AUC, a commonly used evaluation indicator in similar binary classification tasks, is also considered to evaluate the performance of the model. This indicator can comprehensively measure the model's discriminative ability and is particularly suitable for evaluating datasets with unbalanced samples.

The AUC (Area Under Curve, AUC) indicator is defined as the area under the receiver operating characteristic curve (Receiver Operating Characteristic, ROC), which represents the probability value of predicting that the probability of a real image sample is greater than the probability of a forged sample, and its size ranges from 0 to 1. The AUC indicator measures the model's ability to classify real samples and forged samples, that is, the classification performance of the detector: When the AUC value is larger, it means that the detector has a greater probability of ranking positive samples before negative samples, that is, the

classification effect is better. Especially in cross-dataset testing, the AUC indicator can directly reflect the generalization ability of the detector.

5.3 Experimental Results and Analysis

First, experiments are conducted based on the dataset constructed in this paper, and the results are compared with other advanced Deepfake detection algorithms to validate the effectiveness of the proposed approach. Additionally, training is performed on different datasets to conduct cross-dataset validation and comparison experiments, assessing the generalization performance of the algorithm.

To evaluate the proposed method, we first conduct experiments on the FaceForensics++ (FF++) dataset [35], which is divided into training, validation, and test sets in a 6:2:2 ratio. Various evaluation metrics, including AUC (Area Under the Curve), Accuracy, Precision, and Recall, are employed for a comprehensive assessment of the model's performance. Table 2 presents the results across different subsets of FF++.

Table 2: The test performance of intra-dataset on datasets of FF++ [35]

Dataset	AUC (%)	Precision	Recall	Accuracy
DeepFakes	99.75	0.9735	0.9465	0.9741
FaceSwap	99.34	0.9469	0.9867	0.9568
Face2Face	99.10	0.9687	0.9012	0.9589
NeuralTextures	98.45	0.9643	0.9645	0.9489

The proposed method demonstrates exceptional performance on DeepFakes and FaceSwap, achieving AUC values of 99.75% and 99.34%, respectively, indicating a strong capability to detect these types of forgeries. Performance remains consistently high on Face2Face and NeuralTextures, with AUC values of 99.10% and 98.45%, respectively, despite these techniques being more complex than standard face-swapping approaches. As each subset of FF++ contains videos manipulated using a single forgery technique, the introduced artifacts are relatively simple and easier to detect. These findings highlight the effectiveness of the proposed model in identifying forgeries with limited and localized modifications.

To further assess the model's ability to detect more complex deepfake patterns, we conduct additional experiments on the full FaceForensics++ and Celeb-DF datasets. The results, summarized in Table 3, highlight the model's robustness across diverse forgery scenarios. Celeb-DF, in particular, poses a greater challenge due to its wider variety of face manipulations, involving different age groups, genders, and ethnicities. Despite these variations, the proposed model maintains a high AUC of 98.16% and an accuracy of 95.42%, indicating its strong generalization capabilities. On the FaceForensics++ dataset, the model achieves an AUC of 99.24% and an accuracy of 96.28%, confirming its stability in detecting manipulated videos. These results demonstrate that the proposed method is not only effective in detecting specific forgery types but also exhibits strong performance in diverse, real-world deepfake scenarios. The model's ability to generalize across different datasets underscores its robustness and reliability in deepfake detection.

To further validate the effectiveness of our approach, we compare its performance with six state-of-the-art deepfake detection methods, including MesoNet [16], Multi-task [20], Face+Context [21], Multi-attentional [22], M2TR [23], and 3D Residual-in-Dense [37]. The AUC comparisons, presented in Table 4, illustrate the advantages of our model over existing approaches. On the FaceForensics++ dataset, our method achieves an AUC of 99.38%, outperforming M2TR (99.51%) and DINO (99.14%). This demonstrates the effectiveness of incorporating both spatial and temporal information for detecting high-quality

deepfake videos; On the Celeb-DF dataset, our approach achieves an AUC of 98.46%, surpassing M2TR (95.50%) by 2.96% and DINO (96.87%) by 1.59%. These results indicate that our method excels in detecting low-quality and compressed forgeries, where many traditional detection models struggle. Compared to MesoNet and Multi-task models, which rely on shallow feature extraction, our method significantly improves performance, particularly on highly compressed deepfake videos. In contrast to Face+Context and Multi-attentional methods, our approach integrates a Two-Stream 3D Convolutional Network (I3D) and a Vision Transformer (ViT), effectively capturing both temporal inconsistencies and spatial distortions present in deepfake videos.

Table 3: The performance evaluated within the dataset on FF++, Celeb-DF datasets

Dataset	AUC (%)	Precision	Recall	Accuracy
FaceForensics++ [35]	99.24	0.9502	0.9128	0.9628
Celeb-DF [36]	98.16	0.9329	0.9014	0.9542

Table 4: AUC comparisons with 6 state-of-the-art approaches on FF++ and Celeb-DF datasets

Dataset	Precision	FaceForensics++ [35]	Celeb-DF [36]
MesoNet [16]	Video	75.30%	54.80%
Multi-task [20]	Video	80.10%	90.50%
Face+Context [21]	Image/Video Frame	75.00%	66.00%
Multi-attentional [22]	Image/Video Frame	97.60%	_
M2TR [23]	Video	99.51%	95.50%
3D Residual-in-Dense [37]	Video	_	92.93%
DINO [24]	Image/Video Frame	99.14%	96.87%
Ours	Video	99.38%	98.46%

Overall, these comparative results confirm that leveraging both spatial and temporal features is critical for deepfake detection. By integrating I3D for dynamic feature extraction and ViT for spatial context modeling, the proposed method significantly enhances detection accuracy and generalization, particularly in cross-dataset evaluations.

To further evaluate the model's generalization performance, cross-training and testing were conducted on the FF++ and Celeb-DF datasets, with AUC as the evaluation metric. The experimental results presented in Table 5 show that the model achieves high accuracy on known datasets and demonstrates strong generalization performance on unseen datasets. The model trained on Celeb-DF performs less well on FF++, which can be attributed to the greater diversity of facial forgery techniques and generative processes in FF++, making the dataset more varied. Overall, both FF++ and Celeb-DF involve neural network-based face forgery methods, and the model demonstrates reasonable generalization ability. The classification examples of the detection results are shown in Fig. 6.

Table 5: Performance of inter-dataset evaluation on FF++, Celeb-DF datasets

Train	Test	AUC ([37])	AUC ([24])	AUC (Our)
FaceForensics++ [35]	FaceForensics++ [35]	99.57%	98.67%	99.37%
FaceForensics++ [35]	Celeb-DF [36]	68.42%	70.42%	76.10%
Celeb-DF [36]	FaceForensics++ [35]	60.87%	66.21%	73.05%
Celeb-DF [36]	Celeb-DF [36]	98.56%	98.77%	98.63%

Figure 6: Examples of detection results

6 Conclusions

This paper proposes a detection method based on the fusion of spatiotemporal information to learn the spatiotemporal inconsistencies in deepfake videos. The composition and workflow of each module are described in detail. The proposed spatiotemporal feature aggregation network utilizes a dual-branch structure to extract spatial inconsistency features within individual image frames and temporal inconsistency features between consecutive video frames. The Vision Transformer (ViT) module is incorporated to enhance the representation of spatiotemporal inconsistency features, significantly improving detection performance. Experiments on large public datasets, such as FF++ and Celeb-DF, demonstrate that the proposed method enhances detection accuracy. Compared to other advanced detection methods, it also shows improved generalization performance.

7 Limitations and Future Work

7.1 Limitations

Training Speed: The training process is relatively slow due to the need to process each Deepfake video by splitting it into individual frames. This step consumes considerable time. Additionally, the training speed of the ViT network is slower and less efficient. In future work, we plan to focus on extracting representative features from each frame to accelerate the training speed of the network.

Lack of Real-World Face Swap Datasets: Both the FF++ and Celeb-DF datasets primarily contain Deepfake videos created by professional institutions, lacking real-world video data. Furthermore, these datasets predominantly feature faces of foreign individuals, whose facial features may differ from those of people from other regions, making the dataset less representative. A more complete dataset containing real-world face swap data would improve the model's generalization capability.

Accuracy Limitations: While the model has shown improved generalization, accuracy still remains suboptimal. It is currently unable to achieve 100% classification accuracy for faces. As Deepfake technology continues to evolve, detecting Deepfake faces will become increasingly challenging. Some current Deepfake methods are nearly indistinguishable by the human eye, and it is becoming increasingly difficult to identify subtle artifacts. In future research, we will explore the "uniformity" of Deepfake videos to improve detection accuracy and generalization performance.

Challenges in Compression and Low-Light Conditions: Despite its robust performance, the proposed framework encounters limitations in certain challenging scenarios. Specifically, detection accuracy declines when handling videos that have undergone significant compression or are captured under low-light conditions. High compression rates can obscure subtle forgery traces, while poor illumination can degrade the spatial and temporal feature extraction process. These issues underline the need for further optimization.

7.2 Future Work

Future work will focus on improving the framework's robustness against diverse video quality issues, such as compression artifacts and extreme lighting conditions. Techniques like adaptive enhancement for low-light frames and compression artifact removal may help mitigate these challenges. Moreover, incorporating multi-modal data (e.g., audio-visual or metadata) and domain adaptation techniques could improve the model's generalization, enabling more reliable detection across a wider range of real-world scenarios.

Acknowledgement: We sincerely appreciate the valuable comments and suggestions provided by all the reviewers and editors for proofreading and revising this paper. Without their support, we would not have been able to submit this paper in its current form.

Funding Statement: This work was partically supported by National Natural Science Foundation of China (Nos. 62477026, 62177029, 61807020), Humanities and Social Sciences Research Program of the Ministry of Education of China (No. 23YJAZH047), and the Startup Foundation for Introducing Talent of Nanjing University of Posts and Communications under Grant NY222034.

Author Contributions: Conceptualization: Xinyi Wang and Wanru Song; methodology: Xinyi Wang and Chuanyan Hao; software: Xinyi Wang; validation: Xinyi Wang and Feng Liu; writing—original draft preparation: Xinyi Wang and Chuanyan Hao; writing—review and editing, Xinyi Wang, Chuanyan Hao and Feng Liu; funding acquisition: Xinyi Wang, Wanru Song and Chuanyan Hao. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data used in this study are available from FaceForensics++ at https://github.com/ondyari/FaceForensics and Celeb-DF at https://github.com/yuezunli/celeb-deepfakeforensics (accessed on 20 February 2025).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Deepfake. [cited 2019 Oct 29]. Available from: http://github.com/deepfakes/faceswap.
2. Creswell A, White T, Dumoulin V, Arulkumaran K, Sengupta B, Bharath AA. Generative adversarial networks: an overview. IEEE Signal Process Mag. 2018;35(1):53–65. doi:10.1109/MSP.2017.2765202.
3. Kingma DP, Welling M. Auto-encoding variational Bayes. arXiv:1312.6114. 2013.
4. Shi Y, Yang X, Wan Y, Shen X. SemanticStyleGAN: learning compositional generative priors for controllable image synthesis and editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA. p. 11254–64.
5. Wang Y, Chen X, Zhu J, Chu W, Tai Y, Wang C, et al. HifiFace: 3D shape and semantic prior guided high fidelity face swapping. arXiv:2106.09965. 2021.
6. Yang R, Hu X, Huang Z, Zhang Y, Lan R, Deng Z, et al. Review of deep network generative fake face detection methods. J Comput-Aided Des Comput Graph. 2024;36(10):1491–510. doi:10.3724/SP.J.1089.2024.2023-00615.
7. Popescu AC, Farid H. Exposing digital forgeries by detecting traces of resampling. IEEE Trans Signal Process. 2005;53(2):758–67. doi:10.1109/TSP.2004.839932.
8. Matern F, Riess C, Stamminger M. Exploiting visual artifacts to expose deepfakes and face manipulations. In: IEEE Winter Applications of Computer Vision Workshops (WACVW); 2019 Jan 7–11; Waikoloa, HI, USA. p. 83–92.
9. Pishori A, Rollins B, van Houten N, Chatwani N, Uraimov O. Detecting Deepfake videos: an analysis of three techniques. arXiv:2007.08517. 2020.
10. Guarnera L, Giudice O, Battiato S. Deepfake detection by analyzing convolutional traces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops; 2020 Jun 14–19; Seattle, WA, USA. p. 666–7.
11. Ciftci UA, Demir I, Yin L. FakeCatcher: detection of synthetic portrait videos using biological signals. IEEE Trans Pattern Anal Mach Intell. 2020;43(11):1–12.
12. Li Y, Chang MC, Lyu S. In ictu oculi: exposing AI-created fake videos by detecting eye blinking. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS); 2018 Dec 11–13; Hong Kong, China.
13. Li L, Bao J, Zhang T, Yang H, Chen D, Wen F, et al. Face X-ray for more general face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 5001–10.
14. Li L, Zeng WL, Huang YH, Sun WJ. A study on face anti-spoofing based on semi-supervised learning. Chin J Intell Sci Technol. 2021;3(3):370–80.
15. Zhu X, Wang H, Fei H, Lei Z, Li SZ. Face forgery detection by 3D decomposition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 20–25; Nashville, TN, USA. p. 2929–39.
16. Afchar D, Nozick V, Yamagishi J, Echizen I. Mesonet: a compact facial video forgery detection network. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS); 2018 Dec 11–13; Hong Kong, China.
17. Chugh K, Gupta P, Dhall A, Subramanian R. Not made for each other Audio-visual dissonance-based deepfake detection and localization. arXiv:2007.08517. 2020.
18. Qin W, Lu T, Zhang L, Peng S, Wan D. Multi-branch deepfake detection algorithm based on fine-grained features. Computers Mater Contin. 2023;77(1):467–90. doi:10.32604/cmc.2023.042417.
19. Nguyen HH, Yamagishi J, Echizen I. 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); 2019 May 12–17; Brighton, UK. p. 2307–11.

20. Nguyen HH, Fang F, Yamagishi J, Echizen I. Multi-task learning for detecting and segmenting manipulated facial images and videos. In: 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems; 2019 Sep 23–26; Tampa, FL, USA.
21. Nirkin Y, Wolf L, Keller Y, Echizen I. Deepfake detection based on discrepancies between faces and their context. IEEE Trans Pattern Anal Mach Intell. 2021;44(10):6111–21. doi:10.1109/TPAMI.2021.3093446.
22. Zhao H, Zhou W, Chen D, Wei T, Zhang W, Yu N. Multi-attentional deepfake detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 20–25; Nashville, TN, USA. p. 2185–94.
23. Wang J, Wu Z, Ouyang W, Han X, Chen J, Jiang YG, et al. M2tr: multi-modal multi-scale transformers for deepfake detection. In: Proceedings of the 2022 International Conference on Multimedia Retrieval; 2022 Jun 27–30; Newwark, NJ, USA. p. 615–23.
24. Keresh A, Shamoi P. Liveness detection in computer vision: transformer-based self-supervised learning for face anti-spoofing. arXiv:2406.138602024. 2024.
25. Li Y, Lyu S. Exposing deepfake videos by detecting face warping artifacts. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops; 2019 Jun 16–17; Long Beach, CA, USA. p. 46–52.
26. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020.
27. FakeApp. [cited 2025 Jan 1]. Available from: https://www.fakeapp.org/download.
28. Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell. 2017;39(12):2481–95. doi:10.1109/TPAMI.2016.2644615.
29. DeepFaceLab. [cited 2023 Oct 3]. Available from: http://github.com/iperov/DeepFaceLab.
30. DFaker. [cited 2022 May 1]. Available from: http://github.com/dfaker/df.
31. Salvi D, Liu H, Mandelli S, Bestagini P, Zhou W, Zhang W, et al. A robust approach to multimodal deepfake detection. J Imaging. 2023;9(6):122. doi:10.3390/jimaging9060122.
32. Sunanda N, Shailaja K, Babu JS, Krishna KV, LakshmanPratap N, Motupalli RK. Detecting deepfakes: using CNN to identify manipulated visual media. In: 2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC); 2024 Oct 3–5; Kirtipur, Nepal. p. 1428–32. doi:10.1109/I-SMAC61858.2024.10714667.
33. Vaswani A, Shazeer NM, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30. doi:10.48550/arXiv.1706.03762.
34. Carreira J, Zisserma A. Quo Vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017 Jul 21–26; Honolulu, HI, USA. p. 6299–308.
35. Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M. FaceForensics++: learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019 Oct 27–Nov 2; Seoul, Republic of Korea.
36. Li Y, Yang X, Sun P, Qi H, Lyu S. Celeb-DF: a large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 13–19; Seattle, WA, USA. p. 3207–16.
37. Mehra A, Agarwal A, Vatsa M, Singh R. Motion magnified 3-D residual-in-dense network for deepfake detection. IEEE Trans Biom Behav Identity Sci. 2023;5(1):39–52. doi:10.1109/TBIOM.2022.3201887.