# Retraction Notice

At the request of the authors, this paper has been retracted. The authors identified significant errors in the methodology and figures presented in the article. Specifically, they discovered that the facial extraction method described in the paper was incorrectly attributed to dlib, whereas the method used was MTCNN. The misrepresentation affects the validity and reproducibility of the results presented in the study. An inaccuracy in Fig. 2. which illustrates the “Key Frame Extraction” process, depicts the process as being based on the residual relative to the first frame, when in fact, it should be based on the residual relative to the average image of all key frame sequences. The errors fundamentally affect the integrity of the research and its findings. The misrepresentation of the methods and the incorrect figure misled readers and the scientific community regarding the novelty and effectiveness of the methods the authors claimed to have developed. In light of the omissions and accuracies, the authors believe it is their ethical responsibility to retract the paper.

JEI Journal of Electronic Imaging

RESEARCH PAPER

# Exploring spatial–temporal features fusion model for Deepfake video detection

**Jiujiu Wu,[a,b] Jiyu Zhou,[a,c,*] Danyu Wang,[a] and Lin Wang[a]**

[a]Guizhou Education University, School of Physical and Electronic Sciences, Guiyang, China
[b]Universiti Sains Malaysia, School of Computer Science, Pinang, China
[c]Xi'dian University, School of Artificial Intelligence, Xian, China

**ABSTRACT.** The rapid development of Deepfake technology has posed significant challenges in detecting fake videos. In response to the existing problems in reference frame selection, spatial–temporal feature mining, and fusion in face-swapping video detection techniques, we propose a face-swapping video detection model based on spatial–temporal feature fusion. First, key frame sequences are selected using interframe facial edge region differences. Then, the key frame sequences are separately input into the spatial branch to extract hidden artifacts and the temporal branch to extract inconsistent information. Finally, the spatial–temporal features are fused using a self-attention mechanism and input into a classifier to achieve detection results. To validate the effectiveness of the proposed model, we conducted experiments on the Faceforensics++ and Celeb-DF open-source Deepfake datasets. The experimental results demonstrate that the proposed model achieves better detection accuracy and higher-ranking generalization performance than state-of-the-art competitors.

© 2023 SPIE and IS&T [DOI: 10.1117/1.JEI.32.6.063025]

**Keywords:** Deepfake detection; key frame; temporal–spatial fusion; attention block

Paper 230687G received Jun. 8, 2023; revised Nov. 14, 2023; accepted Nov. 27, 2023; published Dec. 8, 2023; retracted Dec. 9, 2024.

## 1 Introduction

Cybersecurity has always been a highly discussed topic. In recent years, the emergence of Deepfake videos have caused significant trouble for information security and social management. In the early stages, creating fake videos required professionals with specialized image processing skills. However, the advent of Deepfake technology has significantly reduced the barrier to creating fake videos. The low-priced cost has led to an explosive growth of online fraudulent videos, thereby exacerbating the threat to network security.

Deepfake face-swapping technology based on deep learning first appeared community forum, which unleashed a wave of Deepfake technology. Since 2018, many open-source Deepfake software or code, such as FaceSwap,[1] Deepfacelab,[2] and FakeApp,[3] have been published on the internet. Such easy-to-use open-source software has led to widespread misuse of Deepfake technology, which not only seriously infringed on the portrait rights of the face-swapping targets,[4] but also caused most people to use the technology to create obscene pornographic videos, challenging the bottom line of the law.[5] Worse still, some people use Deepfake technology to produce images and videos related to political figures, seriously threatening social stability and national security. Therefore, how to detect Deepfake videos is a challenging issue in the field of network security and computer vision.

*Address all correspondence to Jiyu Zhou, zhoujiyu@stu.xidian.edu.cn

1017-9909/2023/$28.00 © 2023 SPIE and IS&T

## 2 Related Works

In recent decades, the FaceSwap technique has gradually moved from the traditional mode of manual feature extraction and machine learning classification to deep learning. More deep neural networks have raised the detection accuracy of fake videos. These deep neural networks can mainly be categorized into three classes based on the video image features utilized by the models: spatial feature-based, spatial–temporal fusion feature-based, and biometric feature-based.

Spatial feature-based models represent a relatively traditional and practical detection approach. They first decompose video into frames and then conduct detection with each frame in different domains. Afchar et al.[6] proposed a lightweight convolutional neural network based on Inception modules to detect forged videos of faces at a mesoscopic level of analysis. Chollet[7] presented an interpretation in convolutional neural networks, considering Inception modules as an intermediate step between regular convolution and depthwise separable convolution operations. In this light, a depthwise separable convolution can be understood as an inception module with a maximally large number of towers. Nguyen et al.[8] designed a network model that combines the visual geometry group network and capsule network to detect face-swapping images. However, these models are vulnerable to the impact of Deepfake technology's spatial feature-based approach, which focus on mining spatial features within individual frames but overlooks interframe features. In contrast, the spatial–temporal fusion feature-based detection method compensates for and integrates the inconsistency between the spatial and temporal dimensions. Li et al.[9] designed a model to expose fake face videos generated with deep neural network models based on detecting eye blinking in the videos, a physiological signal not well presented in the synthesized fake videos. Masi et al.[10] presented a method for Deepfake detection based on a two-branch network structure that isolates digitally manipulated faces by learning to amplify artifacts while suppressing the high-level face content. Zhao et al.[11] proposed a video transformer model based on spatiotemporal self-attention, which detects general Deepfakes by extracting interframe inconsistencies in videos. This method improves the model's performance on unknown forgery by detecting the common inconsistencies in different forgery techniques. Gu et al.[12] delved into the local motion and proposed a novel sampling unit named snippet, which contains a few successive video frames for local temporal inconsistency learning. Moreover, they designed an intra-snippet inconsistency module and an intersnippet interaction module to establish a dynamic inconsistency modeling framework. The module can be embedded in any feature extraction. However, this method adopts a sparse sampling strategy for frames that may be too large to capture the subtle movement inconsistencies between sampling frames caused by motion. The detection based on biometric features is an approach focused on the individual's characteristics, which overcomes the challenges posed by variations in detection algorithm performance due to different forgery techniques and carrier media. Dong et al.[13] proposed using the contrast between a face's internal and external regions as detection features, combined with an external reference dataset, for identity consistency verification. First, they utilized the X-ray method to interchange the internal and external faces of two sets of real images, generating two training datasets. Then, they employed a transformer to extract features from the internal and external regions of the face separately and completed the training process by minimizing the consistency within the internal face and the consistency within the external face. Haliassos et al.[14] designed a RealForensics model, which employs the bootstrap your own latent (BYOL) self-supervised training strategy. Building upon BYOL, RealForensics takes into consideration sound and image modalities. This is specifically manifested by using sound and image as teacher networks separately and learning facial motion representations by leveraging the consistency between the image and audio modalities in real videos.

Despite the emergence of many advanced Deepfake video detection technologies with continuously improving performance, existing detection models still have the following shortcomings: (1) traditional methods randomly sample frames or extract video segments for detection, and the selected information may not be representative, resulting in insufficient representation ability and low model efficiency; (2) most existing spatial–temporal methods' extracted features more tend to mine spatial information, without fully utilizing the correlations between spatial domain and temporal domain. Therefore, in this study, a spatial–temporal features fusion detection model based on key frames is proposed to address the issues above. To verify the effectiveness of this model, experiments were firstly conducted on the FaceForensics++ (FF++) dataset,[15] then selected Celeb-DF as a test sample for cross-dataset detection, aiming to evaluate the model's generalization capability, at last, quantified each proposed module's usefulness by ablation study.

## 3 Methodology

The architecture of spatial–temporal features fusion detection model based on key frames, as shown in Fig. 1, consists of three modules: (1) First, the key frame generation module, which can extract key frames from Deepfake videos as input samples; (2) second, the first frame of the key frames is fed into the spatial branch of the spatial–temporal dual-branch network to extract artifacts information, while the remaining frames are fed into the temporal branch to capture frame-to-frame inconsistencies;[16] and (3) finally, the self-attention fusion and classification module can fuse the spatial–temporal features and evaluate the authenticity of Deepfake videos. The implementation details are as follows.

### 3.1 Key Frame Extraction

As too much redundancy information exists in video, a good sampling approach can improve the model's efficiency while ensuring a certain level of accuracy. In face-swapping videos, helpful information always exists in the facial edge region. Therefore, as shown in Fig. 2, this paper proposes a method for extracting key frame sequences.

First, the image frames were extracted from the Deepfake video, and MTCNN (multitask cascaded convolution neural networks) was used to locate the position of the faces,[17] which is a method that can detect faces through 68 facial landmarks. Furthermore, the face is cropped to form a new frame sequence with a uniform size of $256 \times 256$.

Second, $X$ is poured into two branches. The upper branch in Fig. 2 can calculate the frame difference intensity between each frame and its previous frame, assuming the frame difference is $D_i$. The lower branch can detect 68 facial feature points (only taking the first 27 points) and generate a mask image, namely $M_{\text{mask}}$, to mark each image frame's facial edge region. For more focus on the critical region, our method extracts the facial edge region's interframe difference $D_i'$

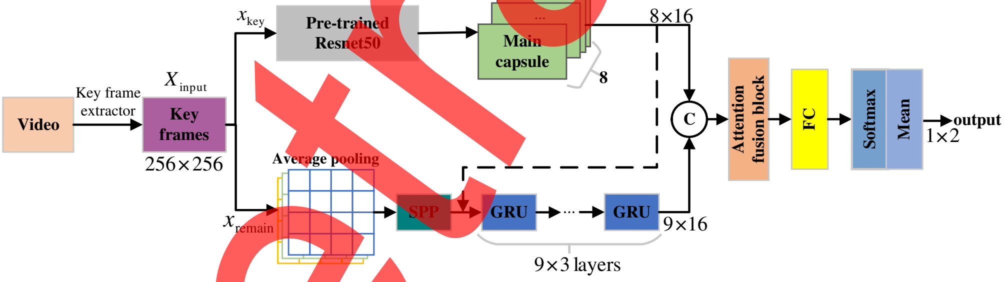

**Fig. 1** The key frame-based spatial–temporal features fusion Deepfake video detection model.

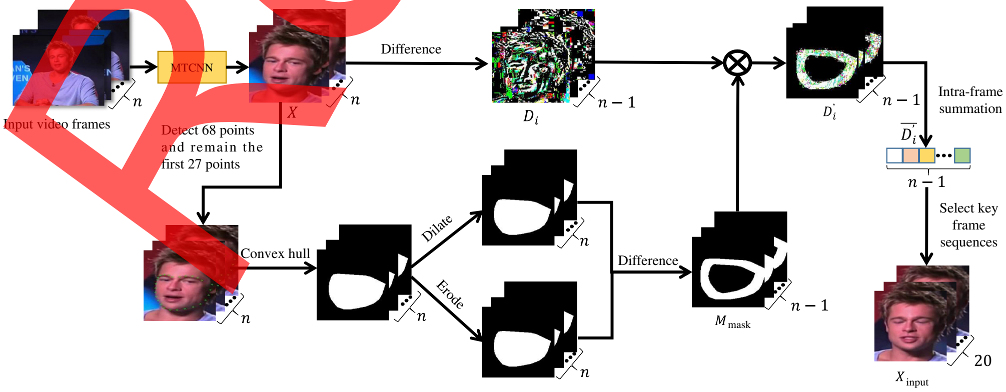

**Fig. 2** Key frame extraction.

by multiplying $D_i$ by $M_{\text{mask}}$ in element-wise. The average intensity of $D_i'$ is defined as $\bar{D}_i'$. The model is described as follows:

$$\bar{D}_i' = \frac{\sum_{x=1}^{w} \sum_{y=1}^{h} D_i'(x, y)}{w \times h}, \tag{1}$$

where $w$ and $h$ denote the width and height of each frame, respectively.

Finally, the key frame owns the greatest $\bar{D}_i$. After the key frame is determined, the following 19 frames are selected to form a video segment along with the key frame, which serves as the input for the detection model

$$\mathrm{X}_{\text{input}} = [x_{\text{key}}, x_{\text{key}+1}, \dots, x_{\text{key}+19}], \tag{2}$$

where $X_{\text{input}}$ represents key frame sequences and $x_{\text{key}}$ indicates the extracted key frame. $[.]$ means cascading operation. The key frame refers to the first frame among the key frames. The subscript "key" means the starting positional index of the key frames within the Deepfake video.

### 3.2 Dual-Branch Spatial–Temporal Network

The dual-branch spatial–temporal network is responsible for extracting important features hidden in the key frames, which reflects the spatial and temporal domain tampering traces left in the Deepfake video during the forgery process. The first branch is the spatial branch, which uses a capsule network[18] to process the key frame extracted through frame difference, i.e., the first frame of the key frames, to learn spatial artifacts. The second branch is the temporal branch, which uses a gated recurrent unit (GRU)[19] to extract temporal features of the remaining key frames and identify frame inconsistencies.

#### 3.2.1 *Spatial branch network*

Extracting features from the face in the spatial dimension can effectively learn the hidden spatial artifacts in the fake face. Since the key frame is the frame with the highest differential intensity selected from the video, it contains the most obvious artifacts. Therefore, the key frame in the key frames was chosen as the input of the spatial branch.

Since the number of frames in a single Deepfake video is limited, it is unsuitable to train the model from scratch. To avoid learning too many high-level details, we use a pre-trained Resnet50 network on the ILSVRC dataset[20] to extract potential features from the before inputting the main frame into the capsule network image. Resnet50 is a typical representative consisting of 50 two-dimensional convolution operations. In the model proposed in this paper, the first to fifth sequence networks of the pre-trained Resnet50 are used: the two blocks in the first convolutional layer and the second convolutional layer. Otherwise, too many convolutional layers will make the network extract high-level semantic information, ignoring the artifact features in the frames, which is not conducive to detection

$$fm_{\text{Resnet50}} = F_{\text{Resnet50}}(x_{\text{key}}). \tag{3}$$

Among them, $F_{\text{Resnet50}}(\cdot)$ represents parts' refer to the Resnet50 network, and $fm_{\text{Resnet50}}$ represents feature maps what were extracted by Resnet50 network.

After extracting the potential features, they are input into the capsule network[21] for spatial feature learning. Generally, the capsule network consists of two parts: main capsules and digit capsules, as shown in Fig. 3. The main capsules are composed of multiple groups of neurons, with each group forming a capsule. Each capsule often has a different structure and can be learned through different feature extraction methods. In this paper, we assigned the same network structure to each capsule to simplify the operation: a 2D convolutional layer, a statistical pooling layer, and a 1D convolutional layer. The statistical pooling layer is used to compute the mean and variance within each convolutional kernel. The main capsules extract the key features, while the digit capsules obtain the final classification results through the attention fusion block. In the spatial branch, we only need to extract the spatial artifact information contained in the key frame, so the main capsules of the capsule network are used for learning key features in the spatial branch

$$f_{\text{cap}}^{i} = F_{\text{capsule}}^{i}(fm_{\text{Resnet50}}), \tag{4}$$

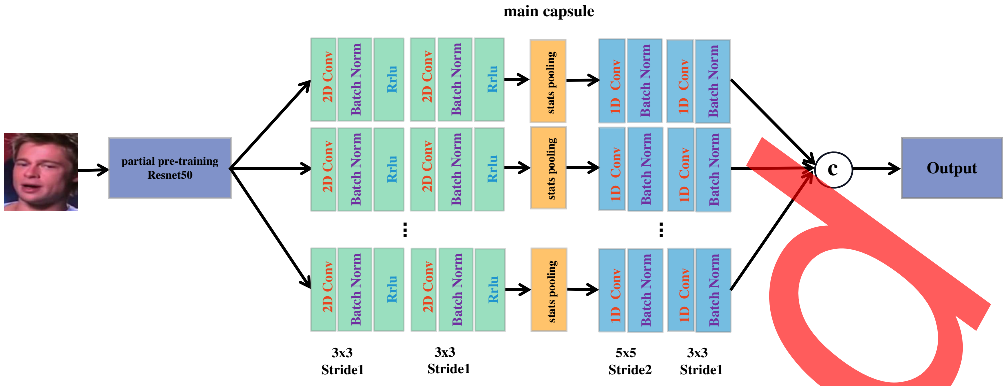

**Fig. 3** Capsule network.

where $f^i_{\text{cap}}$ denotes the features outputted by the $i$'th capsule network, $F^i_{\text{capsule}}(\cdot)$ is the $i$'th capsule

$$f_{\text{cap}} = [fm^1_{\text{cap}}, fm^2_{\text{cap}}, \dots, fm^N_{\text{cap}}]\,, \tag{5}$$

where $N$ represents the number of capsules.

### 3.2.2 *Temporal branch network*

When the first key frame is input into the capsule network for spatial feature extraction, the remaining continuous key frames are input into the temporal branch for extracting temporal correlations. Due to their continuity, key frames are highly similar, which can lead to redundant information and complex calculations. However, the temporal branch detects the temporal inconsistencies between frames, and the spatial information within each frame does not play a key role. Therefore, in the case of extracting spatial artifact information from the key frame, the remaining key frames are differenced from the key frame to obtain a differential image sequence, as shown in the following equation:

$$\Delta x_{i-1} = x_i - x_1 \ \ (i = 2{,}3, \dots, 20)\,, \tag{6}$$

where $x_1$ represents the first frame of $X_{\text{input}}$, $x_i$ represents the $i$'th frame of $X_{\text{input}}$, and $\Delta x_{i-1}$ represents the differential image between the $i$'th frame and the key frame

$$\Delta x = [\Delta x_1, \Delta x_2, \dots, \Delta x_1 9]\,. \tag{7}$$

The differential image is a sparse vector obtained by subtracting the key frames from the first key frame, and directly flattening it would cause spatial waste and greatly increase computational complexity. To address this issue, we used spatial pyramid pooling (SPP)[22] to extract key information from the 3D differential image. As shown in Fig. 4, SPP first performs average pooling on the differential image at different scales and then combines the down-sampled features obtained from each scale into a one-dimensional feature vector as output. This effectively solves the problem of dimension mismatch and avoids resource waste

$$s_i = [F^j_{\text{flatten}}(F^j_{\text{avg}}(\Delta x_i))] \ \ (i = 1{,}2, \dots, 19; j = 1{,}2, \dots, M)\,, \tag{8}$$

where $F^j_{\text{avg}}(\cdot)$ represents the $j$'th average pooling operation before SPP, and $M$ represents the total number of pooling layers, which is usually set between 3 and 5 to reduce significantly the number of parameters. $F^j_{\text{flatten}}(\cdot)$ represents flattening the output of the $j$'th pooling operation into a one-dimensional vector. Finally, concatenate $N$ flattened one-dimensional vectors to obtain the output

$$S = [s_1, s_2, \dots, s_9]\,. \tag{9}$$

So far, a one-dimensional feature vector has been learned from the three-dimensional difference image. The feature vector sequence is then input into a GRU to extract temporal inconsistencies between frames

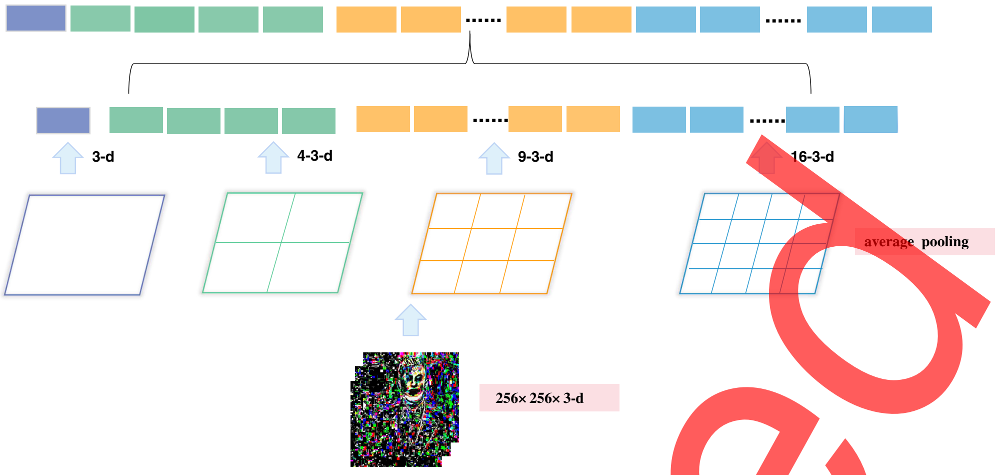

**Fig. 4** SPP.

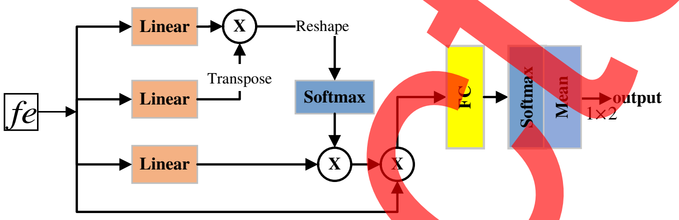

**Fig. 5** Self-attention fusion and classification block.

$$f_{\text{GRUs}} = F_{\text{GRUs}}(S). \tag{10}$$

$F_{\text{GRUs}}(\cdot)$ is $9 \times 3$ layers GRU network in Eq. (10). GRU is a variant of LSTM.[23] Like LSTM, it can solve the problem of vanishing and exploding gradients during long-sequence training. However, GRU computations are more intuitive, which can greatly improve training efficiency. The reason is that GRU can choose to use only one update gate for memorization and forgetting. Compared with LSTM, which requires two update gates, the improvement of GRU greatly reduces the number of parameters and speeds up training.

### 3.3 Classification Based on Self-Attention Fusion

After feature extraction through spatial and temporal flows, the key frames are fully explored for intraframe artifacts and interframe temporal inconsistencies in Deepfake videos. These features are fused as the basis for the next classification.

As shown in Fig. 5, we cascaded spatial and temporal features with the same length of vectors

$$fe = [f_{\text{cap}}, f_{\text{GRUs}}], \tag{11}$$

where $fe$ is then used as an input to the self-attention block

$$\begin{cases} Q = F_Q(fe) \\ K = F_K(fe) \\ V = F_V(fe), \end{cases} \tag{12}$$

where $F_Q(\cdot), F_K(\cdot)$ and $F_V(\cdot)$ mean perform an linear operation on $fe$

$$\hat{y} = \text{softmax}(Fc(\text{softmax}(Q^T K) \cdot V)), \tag{13}$$

where $Fc(\cdot)$ represents full connection operation, and $\hat{y}$ is the final predicted detection result.

## 4 Experiment

### 4.1 Experimental Dataset

To validate the performance and effectiveness of our proposed model, this paper conducted experiments on open-source Deepfake datasets: FF++ and Celeb-DF.[24]

FF++ consists of four types of manipulations: Deepfakes, Face2Face, FaceSwap, and neural textures. FF++ comprises a total of 1000 original videos and 4000 manipulated videos. They are stored at three compression levels: uncompressed, medium compression, and high compression, with compression factors of 0, 23, and 40, respectively. The corresponding video frames have resolutions of 1080p, 720p, and 480p. Detecting the authenticity of uncompressed videos is almost effortless, while medium-compressed videos are relatively easier to handle. However, high-compressed videos present greater challenges due to the blurriness of the frames. In this study, the experiments focused on detecting the authenticity of medium-compressed and high-compressed videos and identifying the four types of manipulations.

Celeb-DF contains 590 original videos collected from YouTube and 5639 corresponding Deepfake videos. The synthesized videos in Celeb-DF exhibit high visual quality, closely resembling the quality of videos circulated online. Therefore, Celeb-DF serves as the ultimate challenge for current Deepfake detection methods. This paper selected Celeb-DF as a test sample for cross-dataset detection, aiming to evaluate the model's generalization capability.

### 4.2 Evaluation Metrics

To evaluate the performance of our proposed spatial–temporal fusion detection model, multiple metrics were selected. First, accuracy is commonly used to evaluate the overall accuracy of a classification model. The higher the accuracy value is, the better the accuracy of the model is. The calculation equation for accuracy $A_{\text{accuracy}}$ is as follows:

$$\mathrm{A}_{\text{accuracy}} = \frac{\mathrm{T}_{\mathrm{TP}}+\mathrm{T}_{\mathrm{TN}}}{\mathrm{T}_{\mathrm{TP}}+\mathrm{F}_{\mathrm{FP}}+\mathrm{T}_{\mathrm{TN}}+\mathrm{F}_{\mathrm{FN}}}. \tag{14}$$

In this equation, $T_{\mathrm{TP}}$ represents the number of correctly classified Deepfake images [true positive (TP)]; $T_{\mathrm{TN}}$ represents the number of correctly classified genuine images [true negative (TN)]; $F_{\mathrm{FP}}$ represents the number of images falsely classified as Deepfake [false positive (FP)]; and $F_{\mathrm{FN}}$ represents the number of images falsely classified as genuine [false negative (FN)].

To comprehensively evaluate the performance of the model, this article used evaluation metrics other than accuracy, including the area under the receiver operating characteristic curve (AUC) of the ROC curve. The ROC curve sorts the samples according to the size of the model's predicted results, takes each sample's predicted probability as a threshold to calculate the FP rate (FPR) and TP rate (TPR) one by one, and plots the curve with FPR as the horizontal axis and TPR as the vertical axis.[25] The formulas for calculating FPR and TPR are as follows:

$$F_{\mathrm{FPR}} = \frac{\mathrm{F}_{\mathrm{FP}}}{\mathrm{F}_{\mathrm{FP}}+\mathrm{T}_{\mathrm{TN}}}, \tag{15}$$

$$T_{\mathrm{TPR}} = \frac{\mathrm{T}_{\mathrm{TP}}}{\mathrm{T}_{\mathrm{TP}}+\mathrm{F}_{\mathrm{FN}}}. \tag{16}$$

The ROC curve can effectively describe the generalization performance of a model. The area under the ROC curve (AUC) is a metric used to evaluate the model's performance, with larger AUC values indicating better performance. The equation for calculating AUC is as follows:

$$A_{\mathrm{AUC}} = \frac{1}{2}\sum_{\mathrm{i}=1}^{m-1}\left(F_{\mathrm{FPR}}^{(i+1)} - F_{\mathrm{FPR}}^{(i)}\right) \times \left(T_{\mathrm{TPR}}^{i} + T_{\mathrm{TPR}}^{(i+1)}\right), \tag{17}$$

where $m$ is the number of examples.

In addition to using key metrics such as accuracy and AUC, this article also employed metrics, such as TP TPR, TN rate, positive predictive value (PPV), and negative predictive value, to evaluate model performance. In the same conditions, a higher value of these metrics indicates better model performance.

### 4.3 Experimental Results and Analysis

The larger the size of the face images cropped from video frames, the more information they contain, leading to more accurate detection results. However, this also leads to higher computational costs. Therefore, in this paper, the image size is set as 256 × 256. This size is sufficient to provide useful information without wasting computational resources. During the training process, batches of 32 videos are used. Two sets of 20 key frames are extracted from each video to form the detection samples for the dual-branch network. Specifically, the frame with the highest frame difference intensity is selected as the key frame for the two sets of key frames, followed by 19 consecutive frames after each key frame. Adam optimization algorithm[26] is employed with a learning rate of 0.0001. All methods are trained with an NVIDA Auadro RTX 3090 GPU (24 GB memory).

This paper compared the proposed model with current Deepfake video detection methods and evaluates the detection performance on the FF++ dataset. Furthermore, the detection effectiveness on the cross-dataset Celeb-DF was further evaluated. The evaluation metrics used in the experiments are ACC (accuracy) and AUC (area under the curve).

#### 4.3.1 *FF++ dataset detection results*

As shown in Fig. 6, demonstrates that as the number of model iterations increases, the classification accuracy of the proposed model gradually improves, indicating its effectiveness. The graph also indicates that the model's accuracy stabilizes around the 60'th iteration, with a final training accuracy of ∼99.03%. This suggests that the proposed model exhibits good classification and detection performance.

On high-quality (HQ) and low-quality (LQ) FF++, the performance of the proposed model in authenticity detection tasks is evaluated, and compared with existing detection methods. As shown in Table 1, the key frame-based spatial–temporal dual-branch detection network almost outperforms existing methods in both ACC and AUC, especially in HQ videos. The proposed method extracts the artifact features from the spatial domain, analyzes the interframe inconsistency of the video from the time domain, and then fuses them through a self-attention fusion block to obtain the best classification prediction, fully mining the inconsistencies contained in the video. The traces of tampering can effectively improve the detection effect.

#### 4.3.2 *Cross-dataset evaluation on Celeb-DF*

It can be seen from the experiment in Sec. 4.3.1 that most of the existing models can reach a higher level of training and verification under a single data set. However, in cross-dataset detection tasks, performance degradation is widespread. To evaluate the generalization performance of

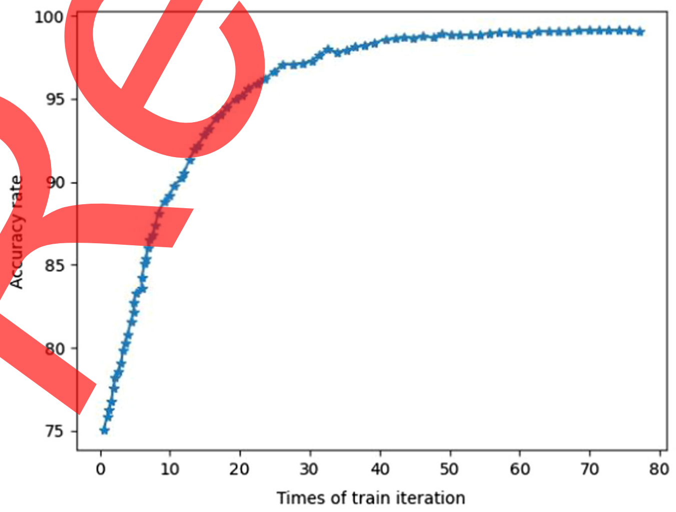

**Fig. 6** Changing curve of accuracy rates with training times.

**Table 1** Comparison of authenticity detection of FF++.

| | FF++ (HQ) | | FF++ (LQ) | |
|---|---|---|---|---|
| Methods | ACC | AUC | ACC | AUC |
| MesoNet[6] | 88.76 | — | 75.65 | — |
| Xception[7] | 95.77 | — | 85.90 | — |
| CNN+LSTM[9] | 96.50 | — | 93.11 | — |
| Capsule network[8] | 96.81 | 97.72 | 93.6 | 95.11 |
| Multi-task[27] | 96.50 | 97.65 | 93.73 | 95.30 |
| Two-branch[10] | 97.43 | 97.90 | 94.10 | 95.96 |
| $F^3$-Net[28] | 99.04 | 99.61 | 94.59 | 96.60 |
| SIM[12] | 98.79 | 99.51 | 94.47 | 96.3 |
| Our model | **99.03** | **99.52** | **94.66** | **96.60** |

Note: bold values indicate the best experimental results under the current evaluation metric.

**Table 2** Comparison of cross-dataset detection effects of Celeb-DF.

| Methods | FF++ | Celeb-DF |
|---|---|---|
| MesoNet | 88.76 | 56.5 |
| Xception | 95.77 | 67.03 |
| CNN+LSTM | 96.50 | 66.4 |
| Capsule Network | 96.81 | 61.1 |
| Multi-task | 96.50 | **75.36** |
| Two-branch | 97.43 | 66.87 |
| $F^3$-Net | 99.04 | 71.93 |
| SIM | 98.79 | 73.41 |
| Our model | **99.03** | 69.18 |

Note: bold values indicate the best experimental results under the current evaluation metric.

the proposed method, this paper first trained the proposed network on HQ FF++ and then tested it on Celeb-DF. Table 2 compares the proposed model and the existing models in detection across datasets. The results show that the training accuracy of the proposed model on FF++ is the highest, and it can also achieve the third-highest performance parameter in the cross-dataset test. Our model still has some gaps with the multi-task and $F^3$-Net in terms of generalization performance, probably because the adopted capsule network pays too much attention to spatially detailed features. But overall, it is still better than most comparable models.

#### 4.3.3 *Ablation study on FF++ dataset*

To verify the effectiveness of the spatial–temporal fusion model, this paper took the key frame extraction module, spatial branch, and classification network as the baseline. Taking the temporal branch and self-attention fusion block as optional modules, the performance of models with different combinations is verified in the FF++ dataset, and the results are shown in Table 3.

It can be seen from Table 3 that the employment of the temporal branch and self-attention greatly improves the accuracy of the model, among which the addition of the temporal branch increases ACC and AUC 0.84% and 1.38%, respectively, and the simultaneous addition of

**Table 3** Ablation study on FF++ dataset.

| Refinements | ACC | AUC |
|---|---|---|
| Baseline | 96.97 | 97.22 |
| + Temporal branch | 97.81 | 98.60 |
| + Temporal branch and self-attention | **99.03** | **99.52** |

Note: bold values indicate the best experimental results under the current evaluation metric.

temporal branch and self-attention module increases ACC and AUC by 2.06% and 2.3%, respectively.

## 5 Conclusion

To improve the efficiency of Deepfake video detection and address the limitations of existing models in input representation and insufficient exploration of spatial–temporal features, we proposed a spatial–temporal features fusion detection model based on key frames. Using interframe facial edge region's differences, we extracted the frame with the highest variations and combined them with consecutive frames to form keyframes, effectively localizing the key frames in the videos. The fusion lately explores the key frames' spatial artifact features and temporal inconsistencies. Experimental results demonstrate the effectiveness of this model, achieving superior detection performance compared to the latest existing methods in the task of authenticating Deepfake videos. Additionally, the proposed model reduces computational complexity compared to existing approaches. However, the model's generalization performance for cross-dataset detection still requires further improvement.

### Disclosures

The authors declare no competing interests.

### Ethics approval

Not applicable.

### Consent to Participate

Written consent for publication was obtained from all participants included in the study.

### Code and Data Availability

The dataset that supports our study will be downloaded from the link https://www.kaggle.com/datasets/sorokin/faceforensics.

### Author Contribution

All authors contributed to the study conception and design.

### Acknowledgments

The authors would like to express their sincere gratitude to all the researchers and mentors involved in this study for their invaluable support throughout the research process. Furthermore, the authors would like to acknowledge the infrastructure provided by Guizhou Education University for this research, as well as support from The Growth Foundation for Young Scientists of the Education Department of Guizhou Province, China, and The Natural Science Foundation of Guizhou Education University, China. Funding: this work was supported by two project grants in China. (1) The Growth Foundation for Young Scientists of Education Department of Guizhou Province, China (Grant No. QJJ [2022] 260). (2) The Natural Science Foundation of Guizhou Education University, China (Grant No. School 2021026).

## References

1. "Github Repository: deep fakes/faceswap," GitHub, 2019, https://github.com/deepfakes/faceswap
2. "Github repository: iperov/DeepFaceLab," GitHub, 2019, https://github.com/
3. "FakeApp," Website, 2019, http://www.fakeapp.com/
4. Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," arXiv:1811.00656 (2018).
5. D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *15th IEEE Int. Conf. Adv. Video and Signal Based Surveill. (AVSS)*, pp. 1–6 (2018).
6. D. Afchar et al., "MesoNet: a compact facial video forgery detection network," in *IEEE Int. Workshop Inf. Forensics and Secur. (WIFS)*, pp. 1–7 (2018).
7. F. Chollet, "Xception: deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 1251–1258 (2017).
8. H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: using capsule networks to detect forged images and videos," in *ICASSP 2019-2019 IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, pp. 2307–2311 (2019).
9. C. M. Liy and L. InIctuOculi, "Exposing AI created fake videos by detecting eye blinking," in *Proc. IEEE Int. Workshop Inf. Forensics and Secur. (WIFS)*, Hong Kong, China, pp. 11–13 (2018).
10. I. Masi et al., "Two-branch recurrent network for isolating deepfakes in videos," *Lect. Notes Comput. Sci.* **12352**, 667–684 (2020).
11. C. Zhao et al., "ISTVT: interpretable spatial-temporal video transformer for deepfake detection," *IEEE Trans. Inf. Forensics Secur.* **18**, 1335–1348 (2023).
12. Z. Gu et al., "Delving into the local: dynamic inconsistency learning for deepfake video detection," *Proc. AAAI Conf. Artif. Intell.* **36**(1), 744–752 (2022).
13. X. Dong et al., "Protecting celebrities from deepfake with identity consistency transformer," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, pp. 9468–9478 (2022).
14. A. Haliassos et al., "Leveraging real talking faces via self-supervision for robust forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, pp. 14950–14962 (2022).
15. A. Rössler et al., "Faceforensics: a large-scale video dataset for forgery detection in human faces," arXiv:1803.09179 (2018).
16. Y. Chen et al., "Deepfake detection with spatio-temporal consistency and attention," in *Int. Conf. Digital Image Comput.: Tech. and Appl. (DICTA)*, IEEE, pp. 1–8 (2022).
17. K. Zhang et al., "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.* **23**(10), 1499–1503 (2016).
18. F. Guidi et al., "High-frame-rate color flow imaging with enhanced spatial resolution in virtual real-time," in *IEEE Int. Ultrasonics Symp. (IUS)*, IEEE, pp. 1–4 (2021).
19. Y. Wu et al., "Enhanced action tubelet detector for spatio-temporal video action detection," in *ICASSP 2020-2020 IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, pp. 2388–2392 (2020).
20. V. U. Prabhu and A. Birhane, "Large image datasets: a pyrrhic win for computer vision?" arXiv:2006.16923 (2020).
21. N. A. Steur and F. Schwenker, "Next-generation neural networks: capsule networks with routing-by-agreement for text classification," *IEEE Access* **9**, 125269–125299 (2021).
22. S. Sriram et al., "Multi-scale learning based malware variant detection using spatial pyramid pooling network," in *IEEE INFOCOM 2020-IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, IEEE, pp. 740–745 (2020).
23. S. Yang, X. Yu, and Y. Zhou, "LSTM and GRU neural network performance comparison study: taking Yelp review dataset as an example," in *Int. Workshop Electron. Commun. and Artif. Intell. (IWECAI)*, IEEE, pp. 98–101 (2020).
24. Y. Li et al., "Celeb-DF: a large-scale challenging dataset for deepfake forensics," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, pp. 3207–3216 (2020).
25. A. Wunderlich, B. Goossens, and C. K. Abbey, "Optimal joint detection and estimation that maximizes ROC-type curves," *IEEE Trans. Med. Imaging* **35**(9), 2164–2173 (2016).
26. I. K. M. Jais, A. R. Ismail, and S. Q. Nisa, "Adam optimization algorithm for wide and deep neural network," *Knowl. Eng. Data Sci.* **2**(1), 41–46 (2019).
27. H. H. Nguyen et al., "Multi-task learning for detecting and segmenting manipulated facial images and videos," in *IEEE 10th Int. Conf. Biom. Theory, Appl. and Syst. (BTAS)*, pp. 1–8 (2019).
28. Y. Qian et al., "Thinking in frequency: face forgery detection by mining frequency-aware clues," *Lect. Notes Comput. Sci.* **12357**, 86–103 (2020).

**Jiujiu Wu** (first author) graduated from Guizhou Education University in 2021 with a bachelor's degree in electronic information science and technology. Currently, she is pursuing a master's degree at the School of Computer Science, Universiti Sains Malaysia. Her current research interests include computer vision and image fusion.

**Jiyu Zhou** (co-first author) is a faculty member at the School of Physics and Electronic Science, Guizhou Education University. He obtained his master's degree in engineering from the Shanghai University of Technology in 2013. Currently, he is a PhD candidate at the School of Artificial Intelligence, Xi'dian University, specializing in computer vision and deep learning. He has published two academic papers, including one in an EI-indexed journal.

**Danyu Wang** received her master's degree in information and communication engineering from the Tiangong University, Tianjin, China, in 2016. Her research interests include nondestructive testing and image processing of partial differential equations. She has published one SCI paper and one Chinese core journals' paper.

**Lin Wang** is currently pursuing her BS degree in information engineering from Guizhou Normal University, Guizhou, China. Her interests include computer vision and deep learning.