Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.057859

ARTICLE

# Lip-Audio Modality Fusion for Deep Forgery Video Detection

**Yong Liu[1,4], Zhiyu Wang[2,*], Shouling Ji[3], Daofu Gong[1,5], Lanxin Cheng[1] and Ruosi Cheng[1]**

[1]College of Cyberspace Security, Information Engineering University, Zhengzhou, 450001, China

[2]Research Institute of Intelligent Networks, Zhejiang Lab, Hangzhou, 311121, China

[3]College of Computer Science and Technology, Zhejiang University, Hangzhou, 310027, China

[4]Henan Key Laboratory of Cyberspace Situation Awareness, Zhengzhou, 450001, China

[5]Key Laboratory of Cyberspace Security, Ministry of Education, Zhengzhou, 450001, China

*Corresponding Author: Zhiyu Wang. Email: wangzhy@zhejianglab.com

Received: 25 August 2024 Accepted: 11 November 2024 Published: 17 February 2025

## ABSTRACT

In response to the problem of traditional methods ignoring audio modality tampering, this study aims to explore an effective deep forgery video detection technique that improves detection precision and reliability by fusing lip images and audio signals. The main method used is lip-audio matching detection technology based on the Siamese neural network, combined with MFCC (Mel Frequency Cepstrum Coefficient) feature extraction of band-pass filters, an improved dual-branch Siamese network structure, and a two-stream network structure design. Firstly, the video stream is preprocessed to extract lip images, and the audio stream is preprocessed to extract MFCC features. Then, these features are processed separately through the two branches of the Siamese network. Finally, the model is trained and optimized through fully connected layers and loss functions. The experimental results show that the testing accuracy of the model in this study on the LRW (Lip Reading in the Wild) dataset reaches 92.3%; the recall rate is 94.3%; the F1 score is 93.3%, significantly better than the results of CNN (Convolutional Neural Networks) and LSTM (Long Short-Term Memory) models. In the validation of multi-resolution image streams, the highest accuracy of dual-resolution image streams reaches 94%. Band-pass filters can effectively improve the signal-to-noise ratio of deep forgery video detection when processing different types of audio signals. The real-time processing performance of the model is also excellent, and it achieves an average score of up to 5 in user research. These data demonstrate that the method proposed in this study can effectively fuse visual and audio information in deep forgery video detection, accurately identify inconsistencies between video and audio, and thus verify the effectiveness of lip-audio modality fusion technology in improving detection performance.

## KEYWORDS

Deep forgery video detection; lip-audio modality fusion; mel frequency cepstrum coefficient; siamese neural network; band-pass filter

## 1 Introduction

In the digital age, deep faking techniques challenge the authenticity of video content. The technology synthesizes or manipulates video through complex algorithms that make it difficult to tell

the real from the fake. Such technological advances not only damage the credibility of the media but also pose risks to social stability and security. Deep forgery can be used to spread disinformation, slander, or influence political outcomes, so effective detection mechanisms are urgently needed. Currently, deep forgery detection mainly relies on visual data analysis, using techniques such as convolutional neural networks to identify anomalies in video frames [1,2]. However, these methods have limited effectiveness when dealing with high-complexity or low-resolution video. Audio analysis, a key part of video authenticity, is often overlooked.

This study proposes a new method that combines visual and audio analysis to detect deepfake videos. Deep learning technology is used based on Siamese neural networks to evaluate the visual authenticity of the video and ensure the consistency of the audio with the lip movement.

The main contributions of the research include: introducing a new feature extraction technology, using MFCC (Mel Frequency Cepstrum Coefficient) and bandpass filter to capture audio features, effectively reducing the masking effect in deep forgery; enhancing Siamese network architecture to improve adaptability to different inputs; proposing a two-stream network structure to process images with different resolutions and improve the detection accuracy. Experiments show that the approach of this study outperforms traditional models in detection accuracy, recall, and F1 scores, and has real-time processing capabilities, which is confirmed through user studies.

Based on summarizing the current research status, this study provides a detailed description of the extraction and processing of lip images, feature extraction, and Siamese network models. Finally, the established theory and methods are experimentally validated using techniques such as multi-resolution image stream experiments, band-pass filter impact testing, and real-time performance testing. The research results confirm the satisfaction and trust level of users towards the detection technology. Through the research in this study, the recognition precision and real-time performance of deep forgery videos can be effectively improved.

The main contribution and innovation of this study lies in the study of a deep forgery video detection technique that integrates lip images and audio signals. The accuracy and robustness of detection are significantly improved through the following key innovative points: firstly, the design of MFCC features combined with band-pass filters is applied, which effectively captures the basic features of audio signals, especially showing advantages in dealing with masking effects; secondly, the Siamese neural network structure is improved by adjusting the architecture and weights of network branches, thus enhancing the adaptability and flexibility of the network to different modal inputs; finally, an innovative two-stream network structure is designed to simultaneously process high-resolution and low-resolution images, fully utilizing multi-resolution information and improving the precision of image matching. The main contribution of this study is to propose a lip sound mode fusion method based on a Siamese neural network for the detection of deepfake videos. By fusing the MFCC features of lip images and audio, combined with band-pass filters to improve audio processing accuracy, the dual-branch Siamese network structure is improved, achieving higher detection accuracy and robustness. Experimental results show that the model has superior performance on the LRW (Lip Reading in the Wild) data set, significantly improving detection accuracy, recall, and real-time processing capabilities.

## 2 Related Work

Currently, many studies focus on the detection of deep forgery videos. Some studies used convolutional neural networks (CNN) to analyze image features in videos to identify forgeries [3–5]. Shende [6] effectively detected deep forgery videos by extracting frame features and analyzing temporal

changes, improving the ability to recognize forgery content. Shelke et al. [7], based on the deep forgery video detection system with VGG-16 (Visual Geometry Group 16) and KPCA (Kernel Principal Component Analysis), implemented forgery detection, effectively addressing the issue of video forgery. Tyagi et al. [8] explored tampered datasets to raise awareness in the research community about privacy and security and promoting the development of universal methods that can detect it. Abhishek [9] classified and located pixel-level forgery, achieving high-precision detection and localization of image forgery. Ganguly et al. [10] designed a deep learning model that combined visual attention technology to distinguish between forgery videos or images and real content. Jin et al. [11] proposed a novel video stitching detection framework, which demonstrated higher F1 scores than existing methods in experiments. Hu et al. [12] proposed a decentralized data training framework for face forgery detection based on federated learning, which applied an inconsistency capture module to enhance the model's ability to recognize dynamic inconsistencies. This framework achieved detection performance comparable to centralized training methods while maintaining data privacy, thereby providing a new solution for protecting information credibility. The above research has summarized various technical methods in the field of deep forgery video detection, aiming to improve the recognition precision and accuracy of forgery content, while emphasizing data privacy protection and model security. However, these methods mainly focus on video modality and ignore the issue of tampering with audio modality.

Siamese neural network is a deep learning model composed of two symmetric subnetworks that share weights and evaluate the similarity between samples by processing input samples in parallel. This architecture effectively quantifies the similarity between samples through embedding representation in high-dimensional space [13–15]. MFCC is used in audio signal processing, especially in speech recognition and deepfake detection, by extracting the spectral features of the audio and enhancing the time-frequency resolution of the speech. Siamese neural networks have many applications in image detection. Kanwal et al. [16] proposed a universal solution based on Siamese neural networks and triplet loss functions for detecting forgery facial images. Ji et al. [17] established a new Siamese framework model that effectively improved intra-frame detection accuracy and inter-frame consistency through self and cross-attention mechanisms. There were also examples of using Siamese neural networks for intrusion detection. Hindy et al. [18] built a single learning model based on Siamese neural networks to enhance the recognition ability of intrusion detection systems for new attack categories. This model learned from limited examples and could recognize unseen attacks without retraining, effectively addressing the shortcomings of traditional supervised learning in terms of data update latency and false positives. Bhatti et al. [19] applied Siamese neural networks for news analysis and proposed an Urdu news story segmentation technique based on long short-term memory-Siamese neural networks. By training the model to recognize positive and negative sentence pairs, transitions between different stories were effectively detected. Madhu et al. [20] applied the deep Siamese capsule network to the one-time detection and classification of malaria thin blood smears. Through feature extraction and discrimination stages, combined with Lorentz similarity measurement, a detection accuracy of 97.24% and a classification accuracy of 98.89% were realized, providing an innovative solution for fast and accurate malaria diagnosis. Rongyu et al. [21] proposed an image-matching method that combined Siamese networks and attention mechanisms, significantly improving the accuracy of image matching and providing a new solution for image similarity detection. Honggui et al. [22] presented a method for identifying discarded mobile phone models based on Siamese convolutional neural networks. The method was used to extract image features through edge detection algorithms and design an adaptive learning rate strategy to optimize model parameter updates, effectively improving recognition precision and speed. Good application results in the sorting of discarded mobile phones were shown. The literature review shows the wide applicability and

effectiveness of Siamese neural networks in a variety of applications, especially for detection tasks. However, these studies have shortcomings in integrating multimodal data, especially in detecting deep-fake videos, where the combination of lip movements and audio signals has not been fully studied. While visual and auditory modes have been extensively studied respectively, their combination potential has not been fully exploited.

In this study, a new deep-forged video detection technology is proposed, which uses a Siamese neural network to perform lip-acoustic mode fusion, to make up for the shortcomings of existing single-peak detection methods and improve the adaptability, efficiency, and accuracy of the detection system. By designing an integrated framework that processes both visual and audio data, this research aims to set a new standard for deepfake video detection, providing a robust solution that can cope with the ever-advancing deepfake technology.

## 3 Lip-Audio Modality Fusion Strategy

### *3.1 Overall Process Design*

In response to the fact that most existing detection methods ignore the issue of audio modality tampering, this study designs a detection process based on the deep forgery video lip-audio matching detection technology using Siamese networks. The specific process is shown in Fig. 1.

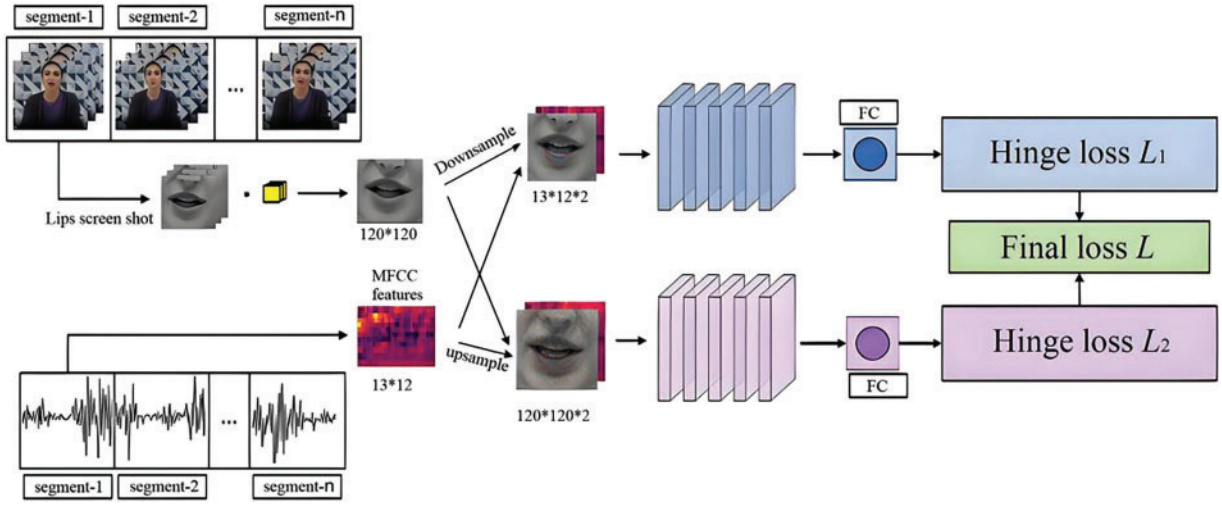


**Figure 1:** Detection process for lip-audio matching in deep forgery videos

In the detection process shown in Fig. 1, first, the input video stream is preprocessed to extract the lip images. These videos are divided into several segments, with lip images extracted in each segment measuring 120 × 120 pixels in size. Next, the audio in the corresponding video segments is preprocessed. MFCC features are extracted, and features with a size of 13 × 12 are generated.

In Siamese networks, video and audio streams are processed by two separate branches. The lip images of the video stream are downsampled to obtain features of 13 × 12 × 2 and input into a branch of the Siamese network. The MFCC features of the audio stream are upsampled to obtain features of 120 × 120 × 2, which are then input into another branch of the Siamese network. After being processed by their respective CNN [23,24], the two branches generate feature vectors, which are then processed by a fully connected (FC) layer to calculate the hinge loss ($L_1$ and $L_2$). The final loss function $L$ is the sum of two hinge losses used to feed back the model parameters to the backend.

### *3.2 Lip Image Processing*

In the process of lip-audio modality fusion, it is necessary to ensure the clarity and consistency of the lip image and to ensure the high quality of the lip image. Firstly, by utilizing technologies such as face detection and feature point localization, accurate extraction of lip shapes from the original video can be achieved, and effective segmentation of lip shapes can be realized. In the facial detection stage, a combination of HOG (Histogram of Oriented Gradients) [25,26] features and linear classifiers is adopted. HOG features capture the directional gradient changes of local object edges in the image, providing a highly robust facial description method that maintains high robustness in changing environments and perspectives. The linear classifier classifies and locates facial regions based on these features, ensuring precise recognition of facial regions.

Once the facial region is determined, the 68-point facial feature point detection model in the Dlib library is used for feature point localization. This model precisely captures the shape and position information of various parts of the face by marking 68 key feature points on the face. Especially feature points 48 to 67, which surround the lips, provide key information for determining the lip bounding box, thus achieving precise segmentation of the lip region. These feature point localization techniques exhibit good stability and accuracy under different lighting conditions, facial expressions, and changes in viewing angles, ensuring the precision of lip area extraction.

Fig. 2 illustrates the process of facial detection technology and feature point localization, as well as extracting the lip region from the overall facial image. The left image displays the detection range of facial feature points. In the image on the right, these key points are numbered, with a total of 68 key points annotated, representing the shape and position information of different parts of the face.

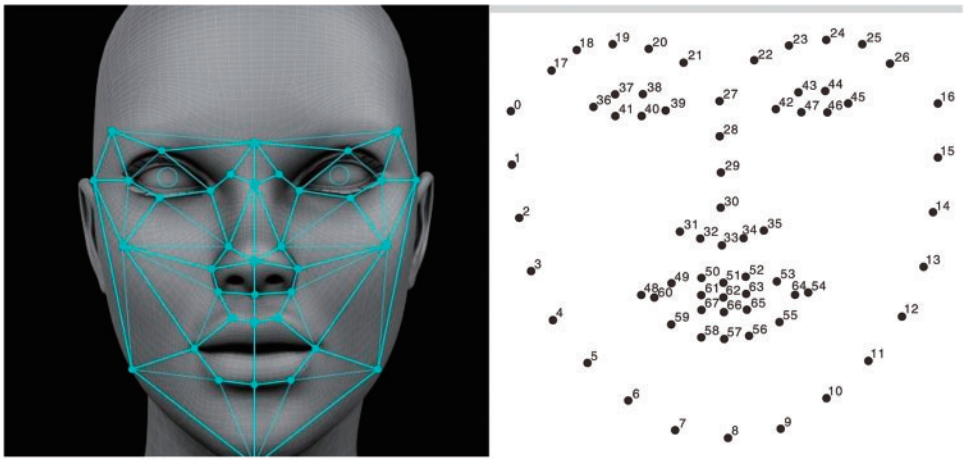

**Figure 2:** Schematic diagram of facial feature point detection and annotation

After extracting the lip area, the next step is to standardize the image. In this process, this study adopts the bicubic interpolation method [27,28], which can preserve the subtle features of the image when scaling and effectively reduce the occurrence of jagged effects. Assuming that it is necessary to find the interpolated value of a certain point in a two-dimensional image, represented as $f(x, y)$, the bicubic interpolation method uses the surrounding 16 points to estimate the value of that point. The calculation formula used is:

$$f(x, y) = \sum_{i=0}^{3} \sum_{j=0}^{3} a_{i,j} \cdot x^i \cdot y^j \tag{1}$$

Among them, $a_{i,j}$ is a coefficient calculated from the surrounding 16 points, and i and j represent the coordinate positioning of the surrounding points. Fig. 3 shows the effect of using the bicubic interpolation method.

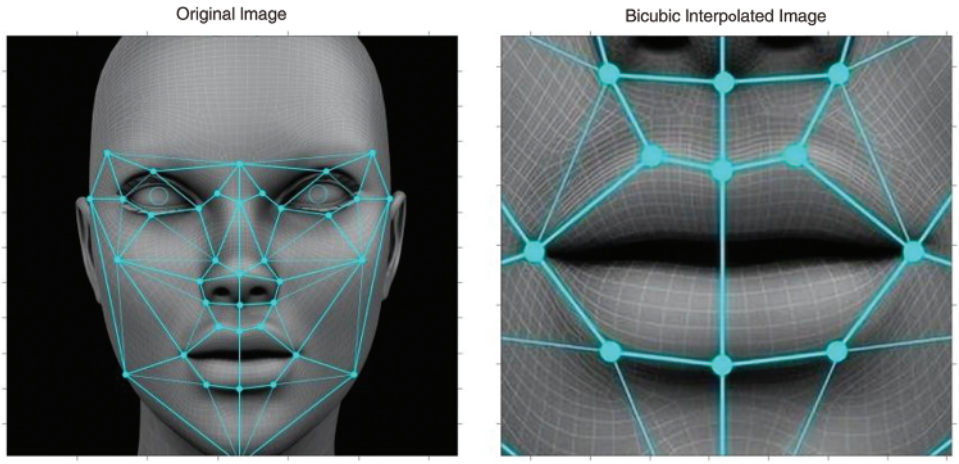


**Figure 3:** Effect of bicubic interpolation method

The left image in Fig. 3 is the original image, and the right image is the enlarged and adjusted image using the bicubic interpolation method. By using the bicubic interpolation method, all lip images are uniformly adjusted to a size of 120 × 120 pixels, which can improve the accuracy and reliability of lip-audio modality fusion in the process of deep forgery video detection, ensure the consistency and alignment of lip regions in each frame of the image, and provide a stable and reliable foundation for subsequent feature extraction and analysis work.

Data augmentation measures are implemented on lip images in the lip-audio modality fusion of deep forgery video detection. The augmentation strategy includes ±10 degree random rotation, ±5 pixel horizontal and vertical translation, 0.9 to 1.1 times scaling, and ±20% brightness adjustment. These augmentation methods implemented through OpenCV significantly improve the diversity of training samples, ensuring lip detection precision under various lighting and poses. To reduce noise in video compression and transmission, the image is smoothed using Gaussian filtering with a standard deviation of 1.5 and a kernel size of 5 × 5 [29,30]. After preprocessing, the features are standardized and aligned. The feature standardization adopts the Z-score standardization method, which achieves the standardization of feature distribution by subtracting the mean of feature values and dividing it by the standard deviation, enhancing the comparability of feature values. The calculation of the mean and the standard deviation is performed on the entire dataset, ensuring consistency between the training and testing phases of the data.

To improve the precision of image alignment, this study adopts an image alignment technique based on affine transformation. By precisely locating the key feature points of the lips, such as the midpoint of the upper and lower lips and the corners of the mouth, and aligning these points to the preset position template, alignment errors caused by facial rotation and tilt are reduced. The affine transformation optimizes the alignment parameters through the least squares method, ensuring the precision and stability of the alignment process.

The technical route is shown in Fig. 4.

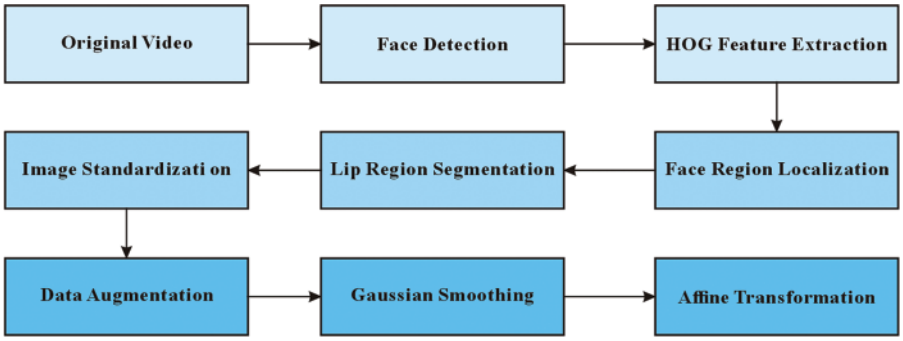


**Figure 4:** Technical route

### *3.3 MFCC Feature Extraction*

MFCC features can effectively capture the spectral features of audio signals, especially the signal details in the presence of masking effects. The process of extracting MFCC features is shown in Fig. 5.

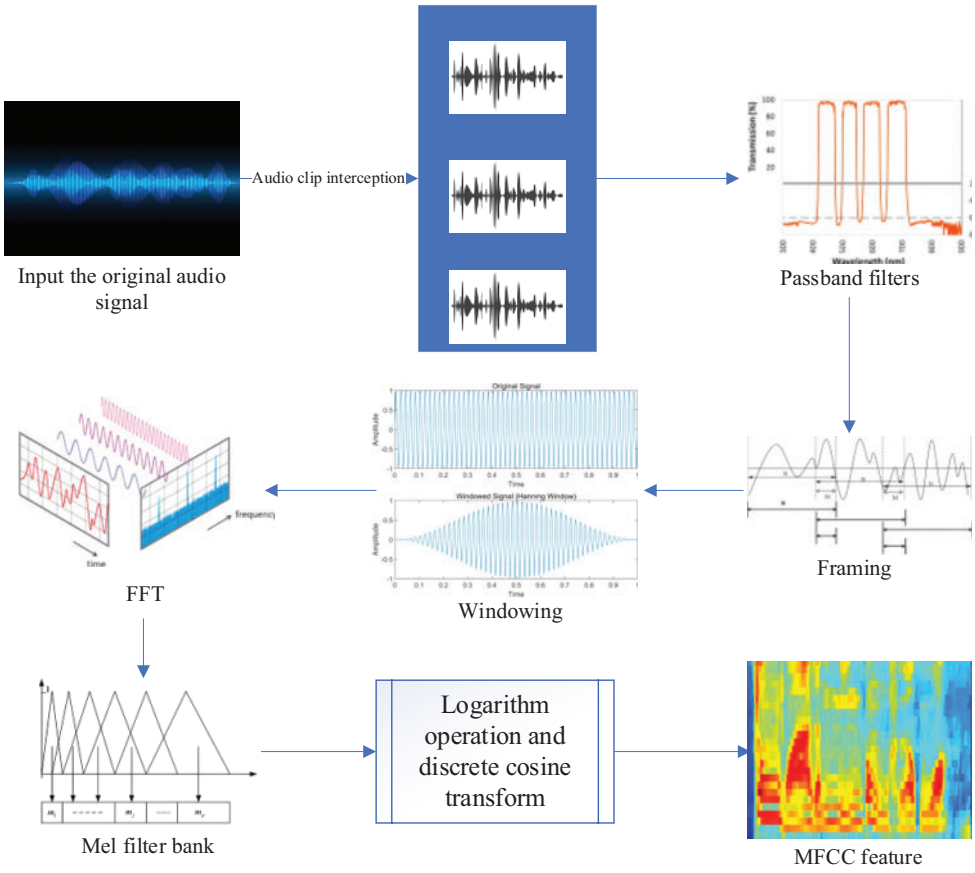


**Figure 5:** MFCC feature extraction

Fig. 5 shows the process of extracting MFCC features. First, the original audio signal is input, and continuous audio signals are divided into smaller frames. This study sets the length of each frame

to 25 ms, and each frame contains 400 sampling points, which can capture the short-term features of the audio signal.

To compensate for the frequency masking effect of speech signals during transmission [31,32] and enhance signal features in different frequency ranges, this study uses a band-pass filter to extract basic audio signal features. The form of a band-pass filter is as follows:

$$H(z) = \frac{z^{-1} - z^{-3}}{1 - 0.9z^{-2}} \tag{2}$$

Among them, H(z) is the transfer function of the band-pass filter; $z^{-1}$ represents the signal being delayed by one cycle in discrete time; $z^{-2}$ and $z^{-3}$ represent delays of two and three cycles, respectively. A band-pass filter allows signals within a specific frequency range to pass through, while blocking signals in other frequency ranges. In the presence of a frequency masking effect, detecting deep forgery videos is more advantageous.

The filtered audio signal is divided into frames of fixed length. Under the condition of a frame length of 25 ms, this study sets the frameshift to 10 ms. Frameshift determines the overlap between adjacent frames, which helps to preserve the temporal information of continuous signals. This overlapping design can make the subsequent processing steps more continuous and smooth in time, avoiding signal breakage caused by abrupt changes between frames.

To perform windowing on audio signals, the windowing function used in this study is the Hanning window, defined as:

$$W(n) = 0.5 - 0.5\cos\left(\frac{2\pi n}{N-1}\right) \tag{3}$$

Among them, n is the sampling point index in the window, and N is the length of the window.

Fast Fourier Transform (FFT) is used to convert time-domain signals into frequency-domain representations [33,34] to obtain the spectral features of the audio frame. It decomposes signals into complex components of different frequencies (including amplitude and phase information) to achieve efficient spectral analysis and precisely identify frequency components in audio. FFT is crucial in speech recognition and audio signal processing, providing the foundation for MFCC feature extraction.

Human auditory features are simulated through a Mel filter bank, and the frequency is converted to Mel scale according to the formula:

$$Mel(f) = 2595\log_{10}\left(1 + \frac{f}{700}\right) \tag{4}$$

Among them, f represents the actual frequency used to describe the vibration frequency of sound, and Mel(f) represents the converted Mel frequency. This conversion relationship reflects that the human ear is more sensitive to low-frequency sounds, while its sensitivity to high-frequency sounds gradually decreases. The application of Mel filters can effectively extract different phoneme frequency components from the speech, which is crucial for deep forgery video detection as it allows for a more accurate capture of acoustic features of the speech. By analyzing Mel frequency, more precise alignment and matching between lip movements and audio content can be achieved, thereby improving the accuracy and reliability of detection.

The logarithmic operation is performed on the spectrum processed by Mel filter bank. By taking the logarithm, the dynamic range of frequency band energy can be effectively compressed, making the eigenvalues smoother and closer to the non-linear perceptual features of human ears towards sound. Subsequently, the discrete cosine transform is applied to further process the energy values after logarithmic transformation. The discrete cosine transform converts these logarithmic energy sequences into cepstral coefficients, which capture the envelope features of the spectrum and are called cepstrum. This study preserves the first 13 cepstral coefficients in the MFCC extraction process, which are the most representative features of key information in audio signals, including intonation and resonance peaks. These coefficients play an important role in lip-audio modality fusion for deep forgery video detection, as they are used to recognize and analyze the consistency between audio and lip movements, providing support for verifying whether the audio and lip movements in the video are synchronized.

### *3.4 Network Architecture Design*

To ensure that the model can effectively integrate and analyze the information of video and audio modalities, this study retains the dual-branch structure of the Siamese neural network and the top-level decision network unchanged in the design. By changing the network architecture composition and weights of the network branches, the flexibility of the network is improved. The Siamese network structure of this study is shown in Fig. 6.

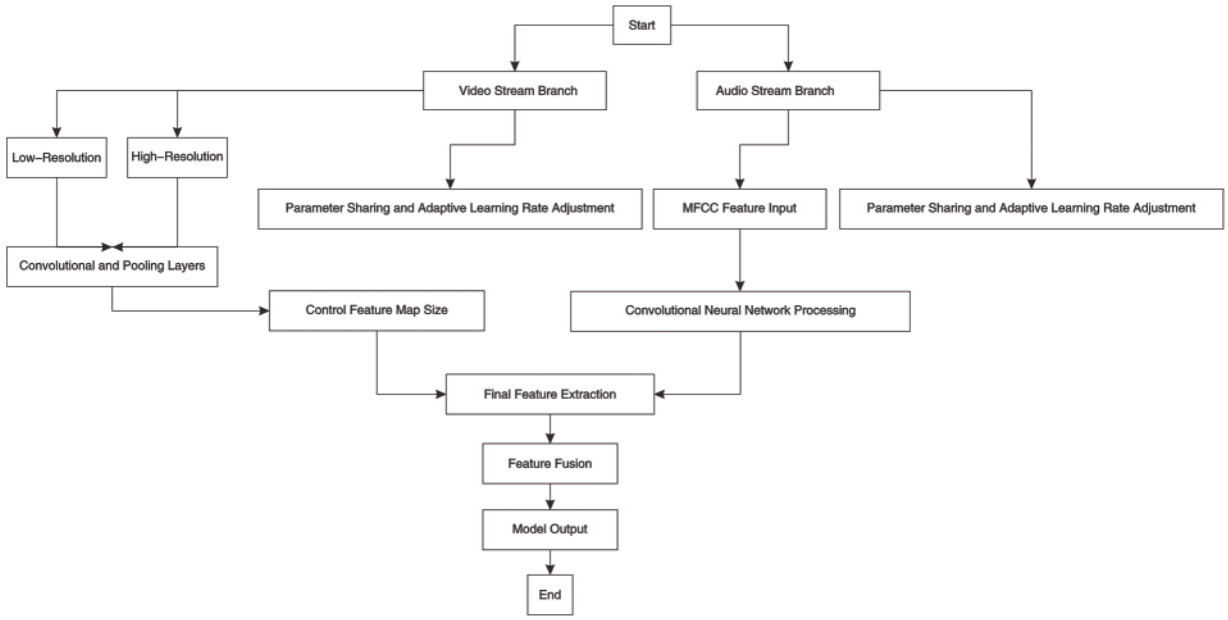


**Figure 6:** Siamese network structure

The Siamese network adopts a dual-branch structure, processing video and audio streams separately to ensure consistency between the two modalities during feature extraction and fusion. Each branch extracts high-level abstract features of video and audio using multi-layer CNN.

In the video stream branch, the lip images of the video stream are processed through multiple convolutional and pooling layers to extract spatial information and feature maps of the images. By gradually reducing the size of the convolution kernel and increasing the stride of the pooling layer,

the input image is reduced from the original resolution of 120 × 120 pixels to the final feature size of 13 × 12, ensuring that necessary information is preserved while reducing computational complexity.

To effectively integrate high-resolution and low-resolution information, this study uses a two-stream network structure to process video streams. The two-stream network structure enhances image-matching precision and robustness by simultaneously processing high and low-resolution inputs. High-resolution streams capture detailed changes, and low-resolution streams process global information. Joint learning improves model adaptability and reduces errors.

In the audio stream branch, MFCC features of 13 coefficients are extracted from preprocessed audio, and advanced features are extracted through convolutional and pooling layers to capture key patterns. To improve network flexibility, parameter sharing and adaptive learning rates are adopted. Parameter sharing reduces the number of model parameters and captures common features of data. Adaptive learning rate adjusts the learning rate based on branch tasks, with audio streams using a smaller learning rate to prevent overfitting and video streams using a larger learning rate to accelerate convergence. Adaptive Gradient Algorithm is used to adaptively adjust the learning rate and optimize specific tasks for each branch.

### *3.5 Loss Calculation and Parameter Update*

After completing the feature extraction of video and audio streams and obtaining feature vectors, the feature vectors of the two branches are further processed through a fully connected layer to generate high-dimensional feature representations. The hinge losses of two branches are calculated to evaluate the degree of matching between video and audio features. If the lip image feature is set to V and the audio feature is set to A, then there are:

$$L1 = \max(0,\ 1 - \text{similarity}(V,\ A_{pos}) + \text{similarity}(V,\ A_{neg})) \tag{5}$$

$$L2 = \max(0,\ 1 - \text{similarity}(A,\ V_{pos}) + \text{similarity}(A,\ V_{neg})) \tag{6}$$

Among them, similarity refers to the measure of similarity between feature vectors; $A_{pos}$ and $A_{neg}$ represent audio features that match or do not match the lip image features, respectively; $V_{pos}$ and $V_{neg}$ represent lip image features that match or do not match audio features. The total loss function L can be obtained by combining the two hinge losses:

$$L = L1 + L2 \tag{7}$$

The total loss function combines the matching error of video and audio features as the optimization objective for model training. By optimizing the total loss function L through backpropagation, the Siamese network gradually optimizes the correspondence between video and audio features, improves the accuracy and robustness of deep forgery detection, and achieves efficient feature extraction and precise lip-sync detection.

## 4 Evaluation of Detection Effect

### *4.1 Datasets*

This study uses the LRW (Lip Reading in the Wild) dataset and the TIMIT (The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus) dataset. LRW includes word pronunciation and lip images, and TIMIT provides multiple pronunciations and detailed audio annotations.

### *4.2 Performance Verification*

To evaluate the effectiveness of lip-audio modality fusion in deep forgery video detection, the LRW dataset is selected for experimental verification in this study. The experiment extracts lip images from video data and MFCC features from audio signals, with a total of 1000 samples selected. The training and testing sets are allocated in a 1:1 ratio. The preprocessed data is then input into various models, including CNN, LSTM, and the model in this study, for training and learning. In addition, to ensure cutting-edge and relevant research, recently proposed advanced methods are used for comparative analysis in the experiments. These models perform lip-audio synchronization matching detection tasks by analyzing the correlation between lip images and audio features. After training, the performance of all models is evaluated using the test dataset. The evaluation results include matching accuracy, precision, recall, and F1 score, as shown in Table 1.

**Table 1:** Performance results of each model

| Model | Type of data | Accuracy (%) | Precision (%) | Recall (%) | F1 score (%) |
|---|---|---|---|---|---|
| CNN [35] | Training set | 85.1 | 84.3 | 88.3 | 86.1 |
| | Test set | 87.5 | 87.4 | 90.2 | 88.2 |
| LSTM [36] | Training set | 83.3 | 85.2 | 84.1 | 85.3 |
| | Test set | 86.4 | 87.3 | 87.3 | 87.2 |
| **The proposed lip-audio modality fusion method** | **Training set** | **88.2** | **88.1** | **91.2** | **89.3** |
| | **Test set** | **92.3** | **91.2** | **94.3** | **93.3** |
| Deep feature frame insertion forgery detection | Training set | 87.4 | 87.3 | 89.4 | 88.4 |
| | Test set | 88.3 | 88.3 | 90.2 | 89.3 |
| Multi-scale convolutional feature fusion forgery detection [37] | Training set | 88.2 | 86.2 | 88.3 | 87.2 |
| | Test set | 91.2 | 89.1 | 89.1 | 89.3 |

Table 1 shows that the accuracy of the model in this study on the test set reaches 92.3%; the precision is 91.2%; the recall is 94.3%; the F1 score is 93.3%. Compared with CNN and LSTM models, the model in this study has advantages in various indicators because it effectively integrates lip image and audio features, and better captures the complex relationship of lip-audio synchronization through deep learning. Especially the high recall rate reflects that the model in this study can recognize more real matches when detecting deep forgery videos, reducing the missed detection rate. In comparison with the recently proposed method, the model in this study is the same on some indexes, but most of the data are better. These results indicate that lip-audio modality fusion has good performance in deep forgery video detection.

### *4.3 Multi-Resolution Image Stream Verification*

The impact of a two-stream network structure on matching precision in deep forgery video detection is verified. Comparative experiments are designed to train Siamese networks using only a

single-resolution image stream and a dual-resolution image stream, respectively, to evaluate the impact of different resolutions on detection accuracy. The video data in the dataset is processed into two formats: high-resolution and low-resolution. The accuracy of the model is tested in three situations: high-resolution, low-resolution, and dual-resolution, and the performance differences are compared. The statistical results are shown in Fig. 7.

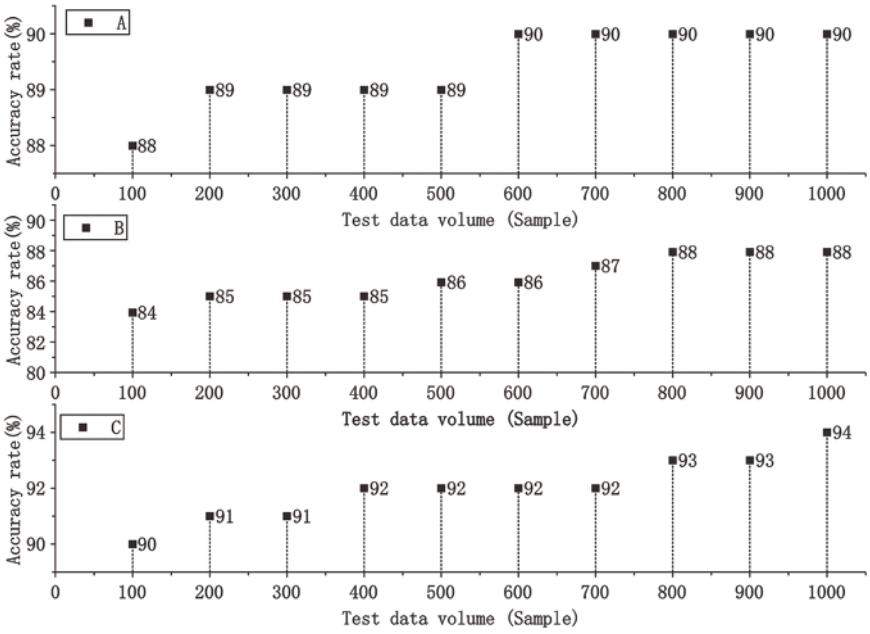


**Figure 7:** Verification results of multiple resolution image streams

In Fig. 7, A–C represent three situations: high-resolution, low-resolution, and dual-resolution, respectively. As the sample size gradually increases, the testing accuracy of the model gradually improves regardless of the resolution of the image. The accuracy of high-resolution image streams gradually increases from 88% to 90%. The accuracy of low-resolution image streams ranges from 84% to 88%. The performance of dual-resolution image streams is better than that of single-resolution image streams at various sample sizes, with a maximum accuracy of 94%. This is because dual resolution can more fully capture and utilize information from different-resolution image streams. The experimental results show that using dual-resolution image streams may be an effective strategy in deep forgery video detection tasks, especially when the sample size increases, its relative advantage becomes more significant, which helps to improve the accuracy of the model.

### *4.4 Test of Effects of Band-Pass Filters*

The impact of band-pass filters on deep forgery video detection when processing different types of audio signals is verified. Models are trained and tested using band-pass filters on processed and unprocessed audio data, and their performance differences in lip-audio modality fusion detection are evaluated. MFCC features are extracted from audio parts using data such as deep forgery videos, real videos, mixed videos, background noise, and multi-person speech as input data. Using SINR (Signal to Interference plus Noise Ratio) and recall as evaluation indicators, the differences between different experimental groups are compared. The statistical results are shown in Table 2.

**Table 2:** Filter test results

| Data type | Experimental group | SINR (dB) | Recall (%) |
|---|---|---|---|
| Deep forgery video | Unfiltered | 18.5 | 85.6 |
| | Band-pass filter processing | 23.7 | 92.3 |
| Real video | Unfiltered | 16.8 | 83.9 |
| | Band-pass filter processing | 21.4 | 90.2 |
| Mixed video | Unfiltered | 17.9 | 84.9 |
| | Band-pass filter processing | 22.1 | 89.1 |
| Background noise | Unfiltered | 15.2 | 81.1 |
| | Band-pass filter processing | 19.8 | 91 |
| Multi-person speech | Unfiltered | 14.7 | 84.8 |
| | Band-pass filter processing | 18.3 | 91.5 |

Table 2 shows the results of training and testing the model using band-pass filters and unfiltered audio data. For different types of data, the SINR processed by band-pass filters is significantly better than that without filtering. In deep forgery videos, the SINR without filtering processing is 18.5 dB, but after band-pass filtering processing, it increases to 23.7 dB, and the recall rate also increases from 85.6% to 92.3%. Other types of data also show similar trends, demonstrating the effectiveness of filters in improving lip-audio modality fusion detection performance. The experimental results demonstrate that band-pass filters reduce the impact of interference and noise, and improve recognition ability.

### *4.5 Test of Real-Time Performance*

The efficiency of the model in real-time video stream processing is evaluated. Video streams containing both real and deep forgery videos are input. After lip image preprocessing and MFCC feature extraction, each frame of video is fed into a Siamese network for matching detection. The number of frames processed per second (FPS, Frames Per Second), average latency, and accuracy are recorded. The results obtained are shown in Table 3.

**Table 3:** Test results of real-time performance

| Video stream type | FPS (frames/s) | Average latency (ms) | Accuracy (%) |
|---|---|---|---|
| Deep forgery video | 30 | 33 | 92 |
| Real video | 25 | 40 | 91 |
| Mixed video | 28 | 36 | 90 |
| High-resolution video stream | 27 | 38 | 93 |
| Low-resolution video stream | 32 | 30 | 90 |

Table 3 shows the performance results of deep forgery video lip-audio matching detection technology based on Siamese networks in different types of real-time video stream processing. When processing deep forgery video streams, the model achieves a processing speed of 30 frames per second, an average latency of 33 ms, and an accuracy rate of 92%. In contrast, the processing speed of real videos is slightly slower, with 25 frames per second and an average latency of 40 ms, but the accuracy

reaches 91%. The highest accuracy is achieved in the high-resolution video stream, reaching 93%. The FPS of the low-resolution video stream is the highest, specifically 32 frames per second because low-resolution video processing is simplified and thus exhibits a faster processing speed. The experimental results show that the proposed lip-audio modality fusion technology has good efficiency and reliability in real-time video stream processing, and is suitable for addressing the detection challenges of various types of videos.

The accuracy and precision of data sets of different sizes are shown in Table 4.

**Table 4:** Accuracy and precision of data sets of different sizes

| Dataset name | Size | Accuracy (%) | Precision (%) |
|---|---|---|---|
| LFW | 13,000 | 92.5 | 89 |
| CelebA | 202,599 | 85.7 | 82.3 |
| VGGFace2 | 3,310,000 | 90.1 | 88.5 |
| UTKFace | 20,000 | 78.4 | 75.6 |
| FER2013 | 35,887 | 88.9 | 86 |

### *4.6 User Research Experiment*

The satisfaction and trust of non-professional users towards the detection results of deep forgery video lip-audio matching detection technology based on Siamese networks are evaluated. 100 non-professional users are selected, and 20 actual cases of model detection are presented. Users evaluate the detection results of the models, and the feedback is collected. They fill out a satisfaction survey questionnaire based on their intuitive perception of the model detection results and rate the trustworthiness of the model's detection results. After collecting data, user satisfaction and trust ratings are analyzed. The satisfaction and trust scores range from 1 to 5, with higher scores indicating greater user satisfaction and trust in the model. The average score of each case is calculated, and the results are shown in Fig. 8.

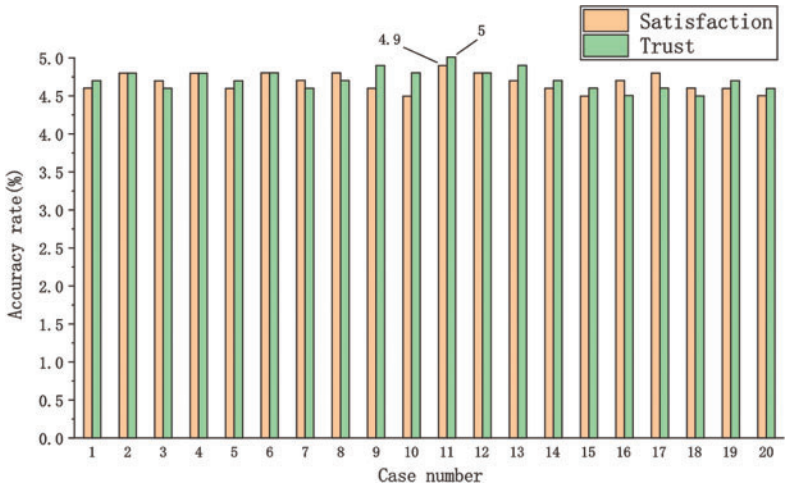


**Figure 8:** Verification results of multiple resolution image streams

Fig. 8 shows the average score statistics of users in each test case. The overall results show that users have a high level of satisfaction and trust in the 20 cases. The average score of Case 11 is the highest, with an average satisfaction score of 4.9 and an average trust score of 5. The average scores of other cases are above 4.5 points. The experimental results show that the test users are satisfied and trust the detection results of this technology, verifying its effectiveness and reliability in practical applications. In deepfake detection, some failure cases are mainly due to insufficient diversity of the data set, which causes the model to fail to effectively recognize complex lip dynamics. Under different lighting conditions or angles, the model's extraction accuracy of lip features is reduced, thus affecting the overall recognition accuracy. In some cases, background noise and video compression lead to a decrease in the synchronization between audio and lip images, which in turn affects the effect of modal fusion. Therefore, increasing the diversity of training data and optimizing feature extraction algorithms are the keys to improving model performance.

## 5 Conclusion

This study studies a deep forgery video detection technology based on Siamese neural networks, which significantly improves the accuracy and robustness of detection by fusing lip images and audio signals. The key innovations include the design that combines MFCC features with band-pass filters, an improved dual-branch Siamese network structure, and a two-stream network structure design. These innovations effectively capture the basic features of audio signals, enhance the adaptability of the network to different modal inputs, and improve the precision of image matching. The experimental results show that the method proposed in this study has excellent performance and receives high ratings in user surveys. The model's reliance on a dual-branch Siamese neural network, coupled with multi-resolution image processing and audio feature extraction, demands significant computational resources. This requirement may hinder its scalability, particularly in resource-constrained environments, limiting its applicability in real-time applications or on devices with limited processing power. Although this study effectively solves the problem of merging lip images with audio, it still has limitations. Future research should consider incorporating facial expressions and body language into the analysis to improve the robustness of deepfake detection. In addition, combining more sensor data and machine learning techniques can help to more comprehensively identify deepfake content, thereby further improving the accuracy of detection.

**Acknowledgement:** This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. I want to acknowledge the invaluable help of my tutor Professor Fenlin Liu, who has given my constant consultant in my paper writing, and has guided me and commented through out the whole process of the papers.

**Funding Statement:** The authors received no specific funding for this study.

**Author Contributions:** The authors confirm contribution to the paper as follows: study conception and design: Yong Liu, Shouling Ji, Zhiyu Wang; data collection: Lanxin Cheng, Ruosi Cheng; analysis and interpretation of results: Yong Liu, Shouling Ji, Daofu Gong; draft manuscript preparation: Yong Liu, Zhiyu Wang. All authors reviewed the results and approved the final version of the manuscript.

**Availability of Data and Materials:** Not applicable.

**Ethics Approval:** Not applicable.

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

## References


[1] S. T. Nabi, M. Kumar, P. Singh, N. Aggarwal, and K. Kumar, "A comprehensive survey of image and video forgery techniques: Variants, challenges, and future directions," *Multimed. Syst.*, vol. 28, no. 3, pp. 939–992, 2022. doi: 10.1007/s00530-021-00873-8.

[2] M. Zanardelli, F. Guerrini, R. Leonardi, and N. Adami, "Image forgery detection: A survey of recent deep-learning approaches," *Multimed. Tools Appl.*, vol. 82, no. 12, pp. 17521–17566, 2023. doi: 10.1007/s11042-022-13797-w.

[3] H. Kaur and N. Jindal, "Deep convolutional neural network for graphics forgery detection in video," *Wirel. Pers. Commun.*, vol. 112, no. 3, pp. 1763–1781, 2020. doi: 10.1007/s11277-020-07126-3.

[4] V. Kumar, M. Gaur, and V. Kansal, "Deep feature based forgery detection in video using parallel convolutional neural network: VFID-Net," *Multimed. Tools Appl.*, vol. 81, no. 29, pp. 42223–42240, 2022. doi: 10.1007/s11042-021-11448-0.

[5] V. Vinolin and M. Sucharitha, "Dual adaptive deep convolutional neural network for video forgery detection in 3D lighting environment," *Vis. Comput.*, vol. 37, no. 8, pp. 2369–2390, 2021. doi: 10.1007/s00371-020-01992-5.

[6] A. Shende, "Using deep learning to detect deepfake videos," *Turk. J. Comput. Math. Educ.*, vol. 12, no. 11, pp. 5012–5017, 2021.

[7] N. A. Shelke and S. S. Kasana, "Multiple forgery detection in digital video with VGG-16-based deep neural network and KPCA," *Multimed. Tools Appl.*, vol. 83, no. 2, pp. 5415–5435, 2024. doi: 10.1007/s11042-023-15561-0.

[8] S. Tyagi and D. Yadav, "A detailed analysis of image and video forgery detection techniques," *Vis. Comput.*, vol. 39, no. 3, pp. 813–833, 2023. doi: 10.1007/s00371-021-02347-4.

[9] J. N. Abhishek, "Copy move and splicing forgery detection using deep convolution neural network, and semantic segmentation," *Multimed. Tools Appl.*, vol. 80, no. 3, pp. 3571–3599, 2021. doi: 10.1007/s11042-020-09816-3.

[10] S. Ganguly, S. Mohiuddin, S. Malakar, E. Cuevas, and R. Sarkar, "Visual attention-based deepfake video forgery detection," *Pattern Anal. Appl.*, vol. 25, no. 4, pp. 981–992, 2022. doi: 10.1007/s10044-022-01083-2.

[11] X. Jin, Z. He, J. Xu, Y. Wang, and Y. Su, "Video splicing detection and localization based on multi-level deep feature fusion and reinforcement learning," *Multimed. Tools Appl.*, vol. 81, no. 28, pp. 40993–41011, 2022. doi: 10.1007/s11042-022-13001-z.

[12] Z. Hu, H. Xie, L. Yu, X. Gao, Z. Shang and Y. Zhang, "Dynamic-aware federated learning for face forgery video detection," *ACM Trans. Intell. Syst. Technol.*, vol. 13, no. 4, pp. 1–25, 2022. doi: 10.1145/3501814.

[13] X. Zhou, W. Liang, S. Shimizu, J. Ma, and Q. Jin, "Siamese neural network based few-shot learning for anomaly detection in industrial cyber-physical systems," *IEEE Trans. Ind. Inform.*, vol. 17, no. 8, pp. 5790–5798, 2020. doi: 10.1109/TII.2020.3047675.

[14] N. An and W. Qi Yan, "Multitarget tracking using Siamese neural networks," *ACM Trans. Multimid. Comput. Commun. Appl.*, vol. 17, no. 2s, pp. 1–16, 2021. doi: 10.1145/3441656.

[15] S. Bharadwaj, S. Prasad, and M. Almekkawy, "An upgraded siamese neural network for motion tracking in ultrasound image sequences," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 68, no. 12, pp. 3515–3527, 2021. doi: 10.1109/TUFFC.2021.3095299.

[16] S. Kanwal, S. Tehsin, and S. Saif, "Exposing AI generated deepfake images using siamese network with triplet loss," *Comput. Inform.*, vol. 41, no. 6, pp. 1541–1562, 2022. doi: 10.31577/cai_2022_6_1541.

[17] Y. Ji, H. Zhang, Z. Jie, L. Ma, and Q. M. J. Wu, "CASNet: A cross-attention siamese network for video salient object detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2676–2690, 2020. doi: 10.1109/TNNLS.2020.3007534.

[18] H. Hindy *et al.*, "Leveraging siamese networks for one-shot intrusion detection model," *J. Intell. Inform. Syst.*, vol. 60, no. 2, pp. 407–436, 2023. doi: 10.1007/s10844-022-00747-z.

[19] M. N. A. Bhatti, I. Siddiqi, and M. Moetesum, "LSTM-based Siamese neural network for Urdu news story segmentation," *Int. J. Doc. Anal. Recognit.*, vol. 26, no. 3, pp. 363–373, 2023. doi: 10.1007/s10032-023-00441-y.

[20] G. Madhu *et al.*, "DSCN-net: A deep Siamese capsule neural network model for automatic diagnosis of malaria parasites detection," *Multimed. Tools Appl.*, vol. 81, no. 23, pp. 34105–34127, 2022. doi: 10.1007/s11042-022-13008-6.

[21] R. Y. Yan, W. Li, Y. M. Chen, H. Huang, W. J. Wang and Y. P. Song, "A Siamese network-based image matching algorithm," (in Chinese), *J. Nanjing Univ. (Nat. Sci.)*, vol. 59, no. 5, pp. 770–776, 2023.

[22] H. G. Han, Q. Zhen, K. Y. Ren, X. L. Wu, Y. P. Du and J. F. Qiao, "Mobile phone model recognition method based on Siamese convolutional neural network," (in Chinese), *J. Beijing Univ. Technol.*, vol. 47, no. 2, pp. 112–119, 2021.

[23] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019, 2021. doi: 10.1109/TNNLS.2021.3084827.

[24] G. W. Lindsay, "Convolutional neural networks as a model of the visual system: Past, present, and future," *J. Cogn. Neurosci.*, vol. 33, no. 10, pp. 2017–2031, 2021. doi: 10.1162/jocn_a_01544.

[25] W. Zhou, S. Gao, L. Zhang, and X. Lou, "Histogram of oriented gradients feature extraction from raw bayer pattern images," *IEEE Trans. Circ. Syst. II*, vol. 67, no. 5, pp. 946–950, 2020. doi: 10.1109/TCSII.2020.2980557.

[26] R. G. Guendel, F. Fioranelli, and A. Yarovoy, "Phase-based classification for arm gesture and gross-motor activities using histogram of oriented gradients," *IEEE Sens. J.*, vol. 21, no. 6, pp. 7918–7927, 2020. doi: 10.1109/JSEN.2020.3044675.

[27] B. K. Triwijoyo and A. Adil, "Analysis of medical image resizing using bicubic interpolation algorithm," *Jurnal Ilmu Komputer*, vol. 14, no. 2, pp. 20–29, 2021. doi: 10.24843/JIK.2021.v14.i01.p03.

[28] Y. Zhu, Y. Dai, K. Han, J. Wang, and J. Hu, "An efficient bicubic interpolation implementation for real-time image processing using hybrid computing," *J. Real Time Image Process.*, vol. 19, no. 6, pp. 1211–1223, 2022. doi: 10.1007/s11554-022-01254-8.

[29] Z. Zhao, T. Karvonen, R. Hostettler, and S. Sarkka, "Taylor moment expansion for continuous-discrete Gaussian filtering," *IEEE Trans. Automat. Contr.*, vol. 66, no. 9, pp. 4460–4467, 2020. doi: 10.1109/TAC.2020.3047367.

[30] B. W. Cheon and N. H. Kim, "Modified Gaussian filter algorithm using quadtree segmentation in AWGN environment," *J. Korea Inst. Inform. Commun. Eng.*, vol. 25, no. 9, pp. 1176–1182, 2021.

[31] L. Mei, Z. X. Xiao, and X. J. Sha, "A method of variable neighborhood search and human ear masking music generation," (in Chinese), *J. Harbin Inst. Technol.*, vol. 52, no. 5, pp. 1–8, 2020.

[32] L. Hui, "Research on sound intensity balance and sound clarity improvement in speaker systems," *Modern Eng. Project Manage.*, vol. 3, no. 7, pp. 115–117, 2024.

[33] C. Eleftheriadis and G. Karakonstantis, "Energy-efficient fast Fourier transform for real-valued applications," *IEEE Trans. Circ. Syst. II*, vol. 69, no. 5, pp. 2458–2462, 2022. doi: 10.1109/TCSII.2022.3163280.

[34] C. G. Dias and L. C. da Silva, "Induction motor speed estimation based on airgap flux measurement using Hilbert transform and fast Fourier transform," *IEEE Sens. J.*, vol. 22, no. 13, pp. 12690–12699, 2022. doi: 10.1109/JSEN.2022.3176085.

[35] S. T. Ikram, S. Chambial, and D. Sood, "A performance enhancement of deepfake video detection through the use of a hybrid CNN deep learning model," *Int. J. Elect. Comput. Eng. Syst.*, vol. 14, no. 2, pp. 169–178, 2023. doi: 10.32985/ijeces.14.2.6.

[36] U. Masud, M. Sadiq, S. Masood, M. Ahmad, and A. A. A. El-Latif, "LW-DeepFakeNet: A lightweight time distributed CNN-LSTM network for real-time DeepFake video detection," *Signal, Image Video Process.*, vol. 17, no. 8, pp. 4029–4037, 2023. doi: 10.1007/s11760-023-02633-9.

[37] A. K. Jaiswal and R. Srivastava, "Detection of copy-move forgery in digital image using multi-scale, multi-stage deep learning model," *Neural Process. Lett.*, vol. 54, no. 1, pp. 75–100, 2022. doi: 10.1007/s11063-021-10620-9.