Exploiting Deepfakes by Analyzing Temporal Feature Inconsistency

Junlin Gu, Yihan Xu, Juan Sun, Weiwei Liu
Jiangsu Vocational College of Electronics and Information,
China

***Abstract*—In recent years, the rapid advancement of image generation technology has facilitated the creation of counterfeit images and videos, posing significant challenges for content authenticity verification. Malefactors can easily extract videos from social networks and generate their own deceptive renditions using state-of-the-art techniques. The latest Deepfake face forgery videos have reached an unprecedented level of sophistication, making it exceptionally difficult to discern signs of manipulation. While several methods have been proposed for detecting fraudulent media, they often target specific aspects, and as new attack methods emerge, these approaches tend to become obsolete. This paper presents a novel detection approach that combines Convolutional Neural Networks (CNN) and Long Short-Term Memory Networks (LSTM). Initially, CNN is employed to extract image features from each frame of the input facial video, capturing subtle alterations and irregularities in manipulated content. Subsequently, the extracted feature sequence is used to train the LSTM network, mimicking the temporal consistency of human visual perception and enhancing the effectiveness of counterfeit video detection. To validate this methodology, a comprehensive evaluation is conducted using the FaceForensic++ dataset, affirming its proficiency in identifying Deepfake counterfeit videos.**

***Keywords*—Face forgery detection; Convolutional Neural Network; Long Short-Term Memory Network; time consistency**

I. Introduction

The rapid and remarkable progress in machine learning technology has elevated the capabilities of video modification and production to an unprecedented level. A pivotal development in this arena is the widespread adoption of Generative Adversarial Networks (GANs), which have revolutionized the automatic generation of images and video synthesis through network model training [1]. Notably, the advent of Deepfake technology, a derivative of GANs, has enabled the seamless replacement of facial features in videos. After post-processing, these videos attain an exceptional degree of realism. However, this rapid technological advancement has brought forth a slew of significant societal challenges [2]. The proliferation of Deepfake technology raises concerns about privacy violations, and its misuse can potentially lead to legal liabilities. Despite diligent efforts by network oversight bodies, the digital landscape remains inundated with a vast volume of synthetic, manipulated videos. It is, therefore, imperative to expeditiously develop effective methods for detecting forged videos to address this burgeoning issue.

As Deepfake technology continues to evolve, the field of detection methods has made substantial progress. Researchers have delved deeply into deep learning models, including spatial domain methods [3], [4], [5] and temporal domain methods [6], [7], in a comprehensive effort to identify irregularities and inconsistencies inherent in Deepfake videos. These models autonomously extract and categorize features, thereby enhancing the accuracy of forged content detection. Moreover, the development and utilization of extensive datasets have played a pivotal role in advancing research on deep forgery detection. Datasets such as FaceForensics++ [8], Deeperforensics [9] have provided a wealth of real and fake video examples, serving as invaluable resources for researchers in this field. The adoption of multi-modal detection approaches, which combine visual data with audio, voice, and other sources of information, has significantly improved the precision and effectiveness of detection methods. Nevertheless, the field of deep forgery detection continues to grapple with an array of challenges. Adversaries consistently refine their Deepfake techniques, making detection increasingly intricate. Consequently, researchers are compelled to continually enhance detection methodologies to bolster their robustness and real-time performance, effectively responding to evolving forgery threats.

In alignment with these advancements, this paper introduces a temporal feature inconsistency analyzing method to enhance the accuracy of Deepfake forgery detection. Specifically, the proposed approach integrates a deep convolutional neural network for image feature extraction and incorporates an LSTM network to analyze correlations between feature sequences. Empirical findings substantiate the efficacy of this methodology, affirming its capacity to facilitate efficient and reliable deep forgery video detection. The main contributions of this paper are as follows:

1) We propose a deepfake detection framework based on the extraction of temporal inconsistencies, combining the feature extraction capabilities of CNN and the temporal feature analysis abilities of LSTM to achieve accurate detection of deepfakes.
2) We tested the algorithm's detection accuracy on video sequences of various lengths using the FaceForensics++ dataset. The experimental results indicate that our algorithm ensures both detection accuracy and computational efficiency when applied to video sequences of 40 frames in length.

This paper focuses on detecting deepfake videos using temporal continuity. Section II provides an overview of recent advancements in deepfake detection research. Section III details the proposed methodology, while Section IV validates its effectiveness through experiments measuring detection accuracy, computational efficiency, and related metrics. Finally, the Section V concludes the paper by summarizing our proposed scheme.

II. Related Work

Currently, research in the field of image forgery detection has made notable advancements. However, there is still a compelling need for further exploration, particularly in the context of real-world scenarios. The domain of forged video detection is predominantly categorized into two main classes: semantic detection and non-semantic detection. The specific detection methods within these categories are systematically organized and illustrated in Fig. 1.

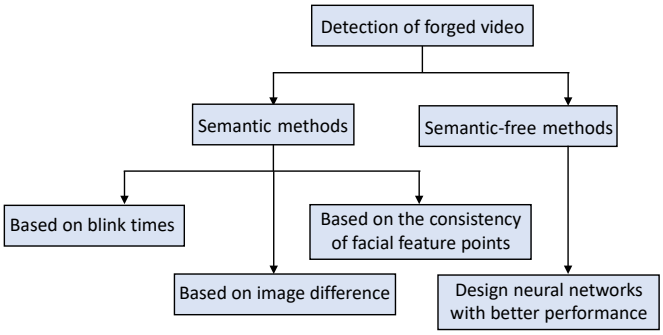

Fig. 1. Classification of forgery video detection methods.

A. Semantic-based Detection Method

1) Classification using the number of blinks in the video: Blinking, denoting the swift opening and closing of eyelids, constitutes a notable behavioral trait. The generation of counterfeit videos through GAN models relies on extensive training data sourced from facial images. As a consequence, many genuine photos do not capture subjects with their eyes closed, leading to a distinctive lack of blink in the generated fake videos. From this vantage point, the absence of blinking emerges as a conspicuous discrepancy between counterfeit and authentic videos.

In the realm of computer vision, blink detection has garnered attention for diverse applications such as fatigue detection [10], [11], [12] and face spoofing detection [13]. Various approaches have been explored in this context. Sukno et al. [14] employed an active shape model in conjunction with optimal invariant features to delineate the eye contour, subsequently assessing eye state based on vertical eye displacement. Torricelli et al. [15] analyzed eye states through the comparison of consecutive frames. Divjak et al. [16] harnessed optical flow to capture eye motion, subsequently extracting the principal eye motion for analysis. Yang et al. [17] deployed parameterized parabolic curves to model eye shapes and fitted the model to individual frames for eyelid tracking.

Drutarovsky et al. [18] delved into the variance of vertical eye area movement as detected by the Viola-Jones algorithm. They further utilized a group of KLT trackers within the eye area, dividing each eye region into 3x3 subregions to calculate the average motion within each. Notably, the most recent development in forged video detection with a focus on blink motion is attributed to Li et al. [19]. Their approach discerns blink occurrences through a combination of Convolutional Neural Networks (CNN) and Long Short-Term Memory Neural Networks (LSTM), ultimately rendering judgments regarding video authenticity based on blink frequencies.

However, this technology primarily hinges on the quantification of blink incidents. Crucially, GAN models employed for the generation of counterfeit videos are trained on a substantial corpus of facial images. In the event that malicious actors augment the training data with closed-eye facial images, the resultant Deepfake counterfeit videos will exhibit plausible blink occurrences, effectively undermining the blink-based detection mechanism.

2) Using the difference between the head pose of the person in the generated video and the head pose of the original video to classify: The face exchange algorithm is designed to generate faces of different individuals while preserving the original facial expressions. However, it is essential to note that the facial feature points of these two faces may not align. The positions of these feature points on the human face are intrinsically linked to crucial structures such as the eyes and mouth. Given that neural network synthesis algorithms cannot guarantee the exact replication of facial features between the original human face and the synthesized face, Yang et al. [20] introduced a novel approach to assess the head pose by comparing estimations derived from all facial feature points with those calculated solely from the central region.

This method is grounded in the observation of errors stemming from the integration of the synthesized face region into the original image. These errors become evident when attempting to estimate the three-dimensional head pose from the facial image. The authors empirically validated this phenomenon through a series of experiments and subsequently devised a classification method based on these observations. It's noteworthy, however, that this approach has yet to be evaluated using the latest Deepfake forged face datasets. Consequently, the question of whether it can effectively detect the most recent Deepfake videos remains an open challenge.

3) Classification by comparing the differences between the face area and the surrounding area: In the realm of image and video detection, recent strides have been taken towards identifying content generated by Generative Adversarial Networks (GANs). Notably, in the context of face exchange, where the original face image from one video is transposed onto the face image of another video, even after a series of fuzzy optimization processes, disparities inevitably emerge between the facial image and its surrounding context.

Li et al. [21] introduced a novel approach that leverages a Convolutional Neural Network (CNN) model to discern discrepancies between the facial region and its neighboring context, thereby facilitating the detection of forged faces. To simulate a broader spectrum of affine distorted faces across different resolutions, the authors trained four CNN models, namely VGG16 [22], ResNet50, ResNet101, and ResNet152 [23]. Subsequently, these models were evaluated by testing them on several synthetic videos sourced from YouTube, affirming the effectiveness of this method for detecting Deepfake forged videos.

B. Non-semantic Detection Method

In recent years, the field of digital image forensics has witnessed a notable integration of deep learning techniques. Rao

and Ni [24] introduced a network dedicated to detecting image stitching, while Rahmouni et al. [25] demonstrated the capacity of deep learning to discriminate between computer-generated and photographic images. These developments underscore the robust performance of deep learning in the domain of digital forensics.

Indeed, traditional microscopic analysis relying on image noise becomes inapplicable within the constraints of compressed video environments, where image noise is often significantly denoised. Similarly, differentiating forged face images at a higher semantic level poses a considerable challenge for the human eye. To address these issues, Darius and Vincent et al. [26] proposed an intermediary approach, employing a deep neural network with a streamlined architecture for image detection. They presented two network structures tailored for detecting forged videos, achieving commendable detection results with a minimal computational overhead. Experimental results showcase an average detection accuracy of 98% for Deepfake counterfeit videos. To further validate the efficacy of this solution, considerable effort was devoted to visualizing the designed network layers and filters.

Another pioneering contribution in the realm of deepfake video detection was presented by Huy et al. [27], who harnessed capsule networks for detecting counterfeit videos across diverse scenarios. This work marked a significant advancement, as it was among the first to explore the application of capsule networks in the field of detection. Capsule networks, initially devised to address issues in digital forensics, were thoroughly examined. The authors conducted a comprehensive analysis and comparison against four mainstream datasets, affirming the superior performance of their method.

However, challenges still persist in the domain of Deepfake video detection, including issues related to the representativeness of datasets and the limited scope of detection. To address these concerns, this paper employs a combination of Convolutional Neural Networks for feature extraction and Long Short-Term Memory networks for analyzing temporal inconsistencies.

III. Problem Analysis and Proposed Method

A. Problems in Deepfake Generation Methods

This section briefly introduces the process of Deepfake generation, and analyzes the problems existing in Deepfake generation method according to its production process.

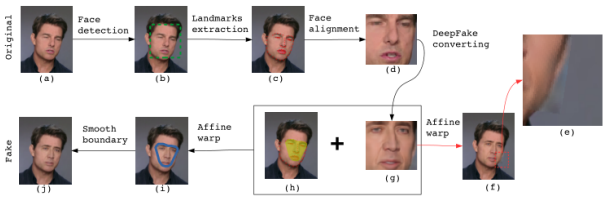

Fig. 2. Forged face generation process diagram.

As illustrated in Fig. 2, the process of generating a forged frame image within a video is detailed. Initially, the original image (a) undergoes face detection to delineate the facial region, depicted as the bounding box in (b). Subsequently, the facial feature points are extracted, and these extracted points are visualized in (c). These feature points convey essential facial characteristics, including facial orientation. Following necessary adjustments, the result in (d) is obtained, which then serves as input to a Generative Adversarial Network (GAN) to produce (g). The subsequent task involves seamlessly integrating (g) with the original image. Two distinct methods are employed for this integration. The first method entails directly replacing (g) with the original image (a) via an affine transformation, generating an image such as (f). However, it becomes evident from (e) that the replaced area does not seamlessly blend with the original image, resulting in noticeable discrepancies. The second approach involves initially identifying the region to be replaced based on the feature points detected in (c), as depicted in (h). This replacement region predominantly corresponds to the central area of the face. Subsequently, (g) is replaced in this region via an affine transformation. Finally, the boundary of the replacement region is softened and smoothed to enhance the image's overall realism. Totally, the problems of Deepfake generation technology are as follows:

1) Intra-frame: When introducing a new face into the target frame image, even with subsequent blurring and smoothing of the boundary, the central facial region tends to exhibit disparities in terms of color, brightness, and resolution when compared to the other areas within the target frame image.

2) Inter-frame: In the context of video face forgery, each image frame undergoes processing, and the GAN employed to generate facial images lacks the ability to retain knowledge of previous frames. In essence, the GAN lacks information about the facial content in the preceding frames, making it challenging to capture the temporal consistency between adjacent frames. Consequently, the facial expressions in consecutive frames may exhibit significant divergence, whereas genuine video sequences tend to maintain a higher degree of consistency in the facial expressions between adjacent frames.

B. The Proposed Method

Building upon David's approach [28], this paper leverages the incongruities present in intra-frame and inter-frame Deepfake video content for detection purposes. In addressing intra-frame disparities, Convolutional Neural Networks are harnessed to extract image features, with the objective of obtaining discriminative features capable of distinguishing genuine from fabricated videos. To address the issue of temporal continuity between frames, this paper adopts the Long Short-Term Memory network (LSTM) for detection. LSTM is a network well-suited for processing sequential data, allowing for the analysis and processing of features that exhibit temporal coherence.

Consequently, this experiment capitalizes on the inconsistency within the image content of Deepfake forged videos and the lack of continuity between adjacent frame images. The process commences with feature extraction performed on each frame within the video, and the resulting feature sequence is subsequently input into the LSTM network. The LSTM network is meticulously trained to identify Deepfake forged videos. The workflow of this solution is visually represented in Fig. 3.

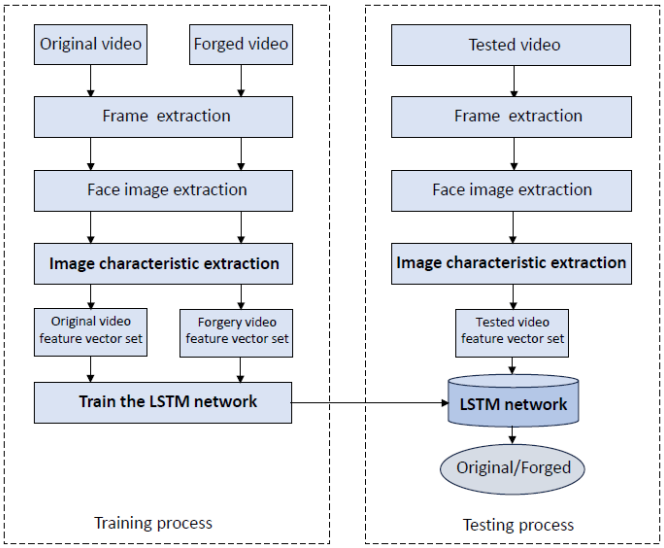

Fig. 3. The workflow of the solution.

1) Video frame extract: Before proceeding with spatial feature extraction, the initial step involves extracting individual frame images from the video footage. To comprehensively evaluate the minimum duration of a forged video that can be reliably and effectively detected, this experiment involves the selection of a sequence of consecutive N-frame images. The study investigates three distinct values for N: 20, 40, and 60 frames, providing a thorough examination of the impact of video length on detection accuracy.

2) Face image extraction: Prior to the spatial feature extraction phase, the fundamental process begins with isolating facial images from the video frames. This initial step holds significant importance for multiple reasons. Firstly, the facial region and its immediate surroundings inherently exhibit a higher degree of distinctiveness, rendering them a valuable asset in the task of differentiating genuine content from manipulated video segments. In our approach, facial detection techniques are employed to pinpoint and delineate prominent facial features within each frame. Once these facial feature points are successfully identified, the corresponding facial images are cropped and separated from the frames. This foundational step lays the groundwork for subsequent spatial feature extraction procedures, ensuring the preservation and utilization of the most informative and distinguishing component of the video - the human face - to enhance the robustness and effectiveness of subsequent detection processes.

3) Image characteristic extraction: Convolutional Neural Networks (CNN) have consistently proven their superiority in image classification tasks due to their exceptional feature extraction capabilities. In the context of forgery video detection, the process of face replacement inevitably introduces inconsistencies between the replaced face image and the surrounding context. These inconsistencies can be effectively captured by the image features extracted through CNN. This paper employs multiple CNNs to individually extract features from face images. The choice of the most suitable neural network model for forgery video detection is determined by comparing their classification performance. These CNNs are utilized to extract image features from video frames within the training and test datasets.

Considering the potential limitations of self-constructed network models in feature extraction, pre-trained network models on the ImageNet dataset, such as VGG19, Inception-V3, and ResNet, are also considered. After removing the last output layer, the output of the final fully connected layer serves as the feature representation for each frame image, thereby facilitating feature extraction from each frame.

4) LSTM network training: To analyzing the temporal inconsistency, the Long Short-Term Memory (LSTM) network is harnessed to analyze the feature sequence extracted by CNN. The LSTM network is equipped with a fully connected layer to map the features derived from LSTM into the ultimate forgery video detection probability. This training process culminates in the development of an LSTM network model designed to serve as a classifier for forged videos.

IV. Experiments

Aiming at the inconsistency in Deepfake forged video, this experiment uses Convolutional Neural Network to extract the spatial features of each frame image to obtain the continuous spatial features of the video, and then inputs the feature sequence into the LSTM network. The sequence features are extracted by the LSTM network to train the classifier, so as to realize the detection of forged video. This chapter introduces the experimental part.

A. Experiment Settings

1) Dataset: This study leverages the widely recognized FaceForensics++ dataset, an extension of the original FaceForensics dataset that has gained extensive adoption in the digital forensics community. The creation of this dataset involved a meticulous process. The dataset production team initiated the project by sourcing 1000 original video files from various online platforms. To ensure the suitability and quality of these videos for research purposes, a minimum resolution of 480p or higher was enforced for each selected video. In recognition of the potential confounding factor of facial occlusion, the production team embarked on a comprehensive manual segmentation process to painstakingly remove any occluded facial fragments within the videos. As a result, the dataset comprises a total of 1000 original videos, with each video yielding an extensive collection of 509,914 individual frames upon the extraction of each image frame. The primary focus of this research lies in the detection of Deepfake counterfeit videos. Therefore, the dataset predominantly draws upon the original video set available in the dataset, complemented by a dedicated Deepfake counterfeit video dataset. The construction and preparation of these datasets were carried out with exceptional care and precision, underscoring their pivotal role in facilitating robust and credible research in the domain of forged video detection. Within the dataset, the training set comprises 720 videos, while the validation and test sets collectively encompass 140 videos. The exact count of video frames included in each dataset can be found in Table I, providing a comprehensive overview of the dataset's

composition. This rigorous dataset design and curation serve as an indispensable foundation, ensuring the availability of a diverse and comprehensive set of video data that is critical for advancing the field of forged video detection.

TABLE I. THE TOTAL NUMBER OF VIDEO FRAMES CONTAINED IN EACH DATASET

Dataset	Training set	Validation set	Testing set
Original	367,282	68,862	73,770
Face2Face	367,282	68,862	73,770
FaceSwap	292,376	54,630	59,672
Deepfakes	367,282	68,862	73,770

2) Experimental parameters: With the network architecture defined, the subsequent step involves data transmission to initiate the training phase. For this experiment, the Adam optimizer, acknowledged for its proficiency in optimizing deep learning models, is employed. The learning rate, a critical hyperparameter influencing convergence and training dynamics, is thoughtfully set to 1e-4 to promote a well-balanced learning process. During the training phase, a designated batch size of 8 examples is input into the network in each training iteration. This iterative process continues until the network achieves convergence, a pivotal juncture in the training cycle where the model has reached its optimal learning capacity. Subsequently, the trained network model undergoes rigorous testing to assess its performance, with a specific focus on detection accuracy. In Table II, provided below for reference, the experiment accounts for the variability in the number of nodes within the fully connected layers in distinct network configurations. To reconcile this variance, systematic adjustments are made to the parameters governing the Long Short-Term Memory (LSTM) layers and the subsequent fully connected layers. These parameter modifications are executed with precision, ensuring the seamless and coherent operation of the network across various architectural configurations.

B. Face Image Extraction

The fundamental premise of this algorithm centers on harnessing discrepancies within the facial region, recognizing that this region exhibits substantial variations indicative of manipulation. Nonetheless, it's essential to acknowledge that a complete video frame contains a plethora of information extending beyond the facial area, which may not be pertinent to the analysis at hand. Therefore, a critical preprocessing step involves the extraction of facial images, with a specific focus on precisely delineating the face and its immediate surroundings. This segmentation process is pivotal in isolating the region of interest, as depicted in Fig. 4, and subsequently refining the dataset for effective analysis. This strategic extraction not only mitigates computational overhead but also streamlines the subsequent processes of feature extraction and classification, thereby enhancing the efficiency and accuracy of Deepfake video detection.

TABLE II. PARAMETER SETTINGS OF THE NETWORKS

Network	VGG19	Inception-V3	ResNet50
LSTM	4096	2048	2048
Fully connected layer	512	512	512

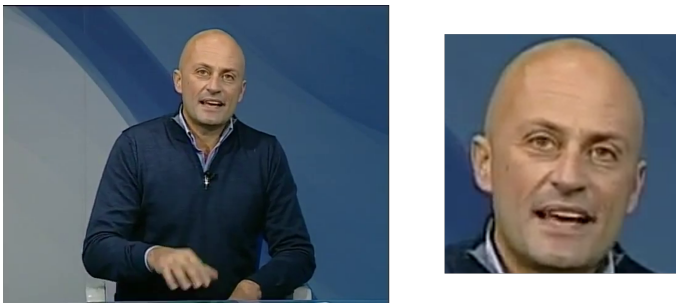

Fig. 4. Comparison diagram before and after face extraction.

C. Image Characteristic Extraction

In this experimental phase, a series of neural networks is employed for spatial feature processing on the dataset. This process is designed to extract one-dimensional features of a fixed length for each image frame extracted from the video sequences. To exemplify this procedure, we will utilize the Inception-V3 network as a representative model. The crux of this operation lies in the extraction of salient features from video frames. The resulting features, pertaining to both unaltered and Deepfake videos, are visually presented in Fig. 5 for reference. The top row showcases the features extracted from a continuous sequence of frames in an unaltered video, while the bottom row illustrates the features extracted from a sequence of frames within a Deepfake video. These visualizations serve to elucidate the distinctions in the extracted features between authentic and manipulated video content. They offer invaluable insights into the characteristic disparities that can be harnessed for Deepfake detection. These visual aids play a pivotal role in comprehending the feature extraction process and its implications for the differentiation between genuine and manipulated video material.

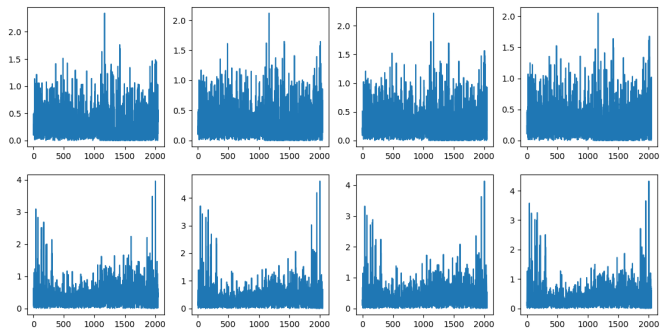

Fig. 5. The extracted feature comparison diagram.

The key observation here pertains to the selected frames extracted from the video, which, being part of a continuous sequence, exhibit an exceptionally high degree of similarity among their image features. This heightened similarity is a direct consequence of the contiguous nature of the frames

within the video. Additionally, it is essential to note that the features extracted from different videos manifest notably distinct characteristics. This pronounced disparity in feature attributes underscores their potential for effective classification.

Fundamentally, these findings emphasize that features extracted from videos possess a discriminative quality, enabling classification to a significant extent. This attribute not only substantiates the feasibility of distinguishing between genuine and manipulated video content but also underscores the effectiveness of feature extraction in augmenting the performance of Deepfake video detection systems.

D. Time Continuity Network Training

In our experiment, which involves the VGG19, Inception-V3, and ResNet networks, it is noteworthy that the output feature sequences generated by these networks exhibit varying lengths. To address this variability, we have meticulously tailored LSTM (Long Short-Term Memory) modules with lengths that correspond to each network's unique feature sequences. This adaptation ensures the effective processing and analysis of the distinct features extracted by each network.

Following the feature analysis conducted by the LSTM modules, the subsequent architectural component comprises a fully connected layer consisting of 512 nodes. This layer plays a pivotal role in consolidating the information derived from the feature sequences. Subsequently, a sigmoid layer is introduced to compute the classification probability. The sigmoid layer serves to transform the output of the fully connected layer into a probability distribution, facilitating the classification of the video content as either authentic or Deepfake.

During the training phase of this experiment, the dataset is divided into a training set, used to train the network model, and a validation set. The network's performance on the validation set is closely monitored, with the classification results offering feedback on the model's effectiveness. Based on the validation outcomes, decisions are made regarding whether to halt the ongoing model training or make further adjustments to model parameters. Iteratively, model parameters are fine-tuned, and this process is reiterated until the optimal model configuration is achieved.

As an illustrative example of the feature extraction process from the ResNet network, the relevant data are visually represented in Fig. 6. This visualization offers insights into the nature of the features extracted from the ResNet network and their potential to enhance the Deepfake detection process.

The analysis of the four curves provides valuable insights into the performance of the LSTM network in our experiment. It is evident that the LSTM network demonstrates a commendable ability to effectively fit the training set, producing outcomes that closely align with the ground truth labels. However, when the same network is applied to the validation set, the results appear to be comparatively less accurate. This performance discrepancy between the training and validation sets reveals a couple of key observations. Firstly, this observed difference underscores the LSTM network's capability to effectively harness the features extracted by the Convolutional Neural Network (CNN) for classification purposes. The capacity of the LSTM network to adapt to

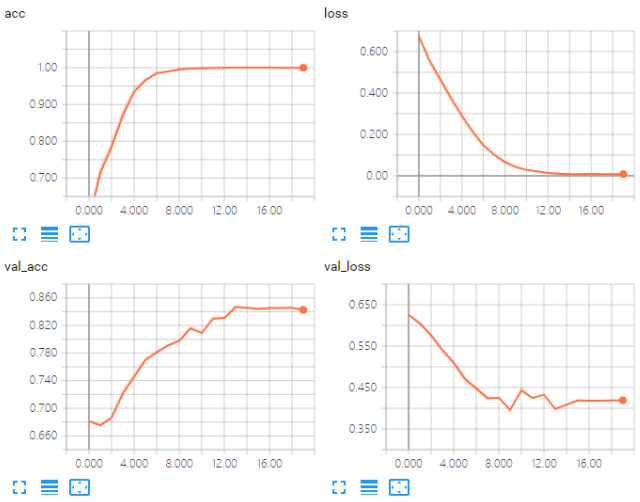

Fig. 6. Network training curve diagram.

and learn from the features derived from the CNN highlights the symbiotic relationship between these components within the deep learning framework. On the other hand, the noted performance gap between the training and validation sets also suggests that the dataset size may be insufficient to achieve a perfect fit to the validation set. This situation is not uncommon in the field of machine learning, particularly when the model tends to memorize the training data rather than generalize to unseen data. Therefore, the results underscore the necessity for larger and more diverse datasets to bolster the network's performance on validation data, thereby enhancing its ability to make accurate classifications in real-world scenarios.

E. Detection Accuracy

In our experimental design, we deliberately limited our analysis to three specific lengths of consecutive video frames: N = 20, 40, and 60. The rationale behind conducting these three distinct sets of experiments was to ascertain the minimum video length necessary for effective Deepfake video detection. To accomplish this, we harnessed the capabilities of four distinct networks for spatial feature extraction. Subsequently, we employed LSTM (Long Short-Term Memory) for the analysis of sequence features, ultimately culminating in the determination of classification accuracy for detecting forged video, as exemplified in Table III.

Analyzing the results depicted in Fig. 7, it becomes evident that ResNet has consistently demonstrated outstanding performance across the various video clip lengths used in the experiments. This observation underscores that the features extracted by ResNet are notably representative and adaptable in the context of classification challenges like video forgery detection. In contrast, the classification results of the simpler Convolutional Neural Network (CNN) on the test set appear to be less impressive. The primary reason for this disparity lies in the absence of pre-training using large-scale datasets. As a consequence, the features extracted by the simple CNN do

TABLE III. CLASSIFIED RESULTS

Number of video frames	Simple CNN + LSTM	VGG19+LSTM	Inception-V3+LSTM	ResNet+LSTM
20	0.5096	0.7749	0.762	0.7829
40	0.5113	0.7781	0.7669	0.8327
60	0.5112	0.8039	0.7572	0.8472

TABLE IV. TIME CONSUMPTION OF RESNET+LSTM

Video frame length	Frame extraction(s)	Face extraction(s)	Feature extraction(s)	Classification(s)
20	18.26	11.88	79	56
40	19.52	22.83	148	94
60	19.97	34.1	222	135

not effectively capture the distinctions between different video frames.

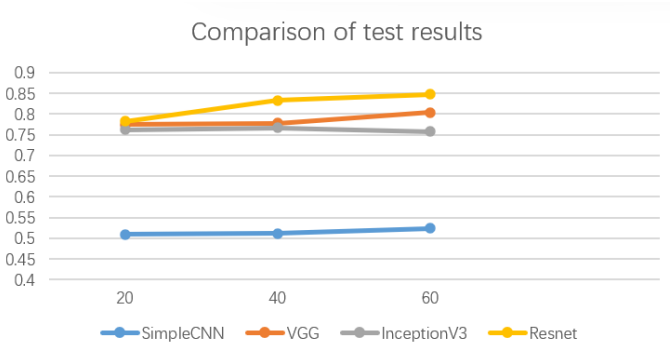

Fig. 7. Classification accuracy comparison chart.

Furthermore, it is noteworthy that there are variations in detection accuracy among video clips of differing lengths. These differences arise due to variations in the features extracted by different networks, with these variations impacting the representation of features concerning temporal continuity. As the length of the video clip increases, the classification accuracy gradually improves. This trend suggests that in longer videos, the contiguous frames more effectively reflect the temporal continuity. For instance, taking ResNet as an illustration, a video clip with a length of 40 frames attains a commendable 83.27% detection accuracy. Expanding the length to 60 frames yields only a modest 1.5% improvement in accuracy, while also increasing the computational complexity. As a result, in practical applications, opting for 40-frame video clips for detection represents a reasonable compromise. Similar trends can be observed in the performance of other networks in the experiments.

F. Time Consumption

Focusing on the ResNet network's performance within the experiment, this paper utilizes the ResNet architecture in combination with LSTM for a more detailed analysis of the method's time consumption.

Table IV illustrates the time allocation for various stages of the method, revealing that the processes of frame extraction and face extraction consume a considerable amount of time. The duration of these processes is also influenced by the system's hardware performance. As a remedy, when implementing this algorithm within a system, these two time-consuming steps can be executed in the background to mitigate user wait times and enhance system efficiency.

Guided by the comprehensive analysis of classification accuracy and time consumption across the experiment, the results point to the utility of 40-frame video clips. This length not only yields superior classification accuracy but also minimizes the computational overhead, resulting in a more efficient and responsive system.

G. A Comparison with Existing Researches

The analysis of the experimental results highlights the effectiveness of the methodology that utilizes Convolutional Neural Networks (CNN) for feature extraction, complemented by Long Short-Term Memory (LSTM) networks for sequence feature extraction. Particularly, the features extracted by ResNet prove to be the most suitable for the task of Deepfake video detection.

Table V offers a comprehensive overview of performance metrics for various algorithms, as provided by the FaceForensic++ dataset production team, with a specific emphasis on Deepfake video classification. Each method in the table is labeled with 'c23' and 'c40,' denoting the video compression rate used. A notable observation from the table is the highest reported detection accuracy of 0.882, as provided by the dataset production team. In contrast, the best result achieved in our experiment stands at 0.924, surpassing the performance of other detection techniques listed in the table.

This outcome serves as robust validation of the effectiveness of the methodology we have employed. The method capitalizes on CNNs for image feature extraction and subsequently subjects these feature sequences to LSTM network training and testing. This comprehensive approach demonstrates the method's ability to effectively detect Deepfake forged videos, as evidenced by its superior performance relative to other techniques in the comparative analysis.

TABLE V. FACEFORENSIC++ DATASET DETECTION ACCURACY TABLE OF EACH METHOD

Methods	Detection accuracy(%)
Ours	**0.924**
Bayar c23 [29]	0.882
Recast c23 [30]	0.836
XceptionFull(FaceForensics++) [31]	0.755
MesoNet c40 [26]	0.700
Rahmouni c40 [25]	0.691

V. CONCLUSION AND FORESIGHT

This paper addresses the challenge of forged video detection as a binary classification task and introduces a novel approach that leverages the synergy between Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM), yielding outstanding classification results. CNN has demonstrated its prowess in computer vision, affirming its exceptional feature extraction capabilities. The methodology effectively harnesses CNN for comprehensive feature extraction on individual image frames, followed by an in-depth analysis of these feature sequences using LSTM, culminating in reliable forged video detection. The experiment was conducted using the FaceForensics++ dataset, encompassing a substantial number of manipulated videos. The results unequivocally demonstrate the remarkable efficacy of our method in detecting Deepfake face forgery videos. In our future research endeavors, we are committed to ongoing enhancements of existing classification algorithms to further elevate the accuracy of Deepfake face forgery video detection. We are enthusiastic about the evolving landscape of this research domain and firmly believe that these advancements will yield more reliable and efficient solutions for forged video detection.

ACKNOWLEDGMENT

This work is supported in part by the Jiangsu Province Department of Industry and Information Technology Key Technology Innovation Project Orientation Program under grant numbers 141-62-65, in part by the Jiangsu Provincial Science and Technology Department Digital Public Service Platform Project under grant numbers 93208000931, in part by the Jiangsu Provincial Department of Science and Technology Industry-Academia-Research Project under grant numbers BY20221343, in part by Jiangsu Provincial Vocational Education Big Data Technology 'Double-Teacher' Master Studio Project, in part by the Jiangsu Province large-scale scientific instrument open sharing project under grant numbers TC2023A073.

REFERENCES

[1] A. Kammoun, R. Slama, H. Tabia, T. Ouni, and M. Abid, "Generative adversarial networks for face generation: A survey," *ACM Computing Surveys*, vol. 55, no. 5, pp. 1–37, 2022.

[2] M. Westerlund, "The emergence of deepfake technology: A review," *Technology innovation management review*, vol. 9, no. 11, 2019.

[3] L. Guarnera, O. Giudice, and S. Battiato, "Deepfake detection by analyzing convolutional traces," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 666–667.

[4] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, "Learning self-consistency for deepfake detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 023–15 033.

[5] L. Chen, Y. Zhang, Y. Song, J. Wang, and L. Liu, "Ost: Improving generalization of deepfake detection via one-shot test-time training," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 597–24 610, 2022.

[6] J. Guan, H. Zhou, Z. Hong, E. Ding, J. Wang, C. Quan, and Y. Zhao, "Delving into sequential patches for deepfake detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 4517–4530, 2022.

[7] D. M. Montserrat, H. Hao, S. K. Yarlagadda, S. Baireddy, R. Shao, J. Horváth, E. Bartusiak, J. Yang, D. Guera, F. Zhu *et al.*, "Deepfakes detection with automatic face weighting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 668–669.

[8] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1–11.

[9] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2889–2898.

[10] M. Kołodziej, P. Tarnowski, D. J. Sawicki, A. Majkowski, R. J. Rak, A. Bala, and A. Pluta, "Fatigue detection caused by office work with the use of eog signal," *IEEE Sensors Journal*, vol. 20, no. 24, pp. 15 213–15 223, 2020.

[11] A. Kuwahara, K. Nishikawa, R. Hirakawa, H. Kawano, and Y. Nakatoh, "Eye fatigue estimation using blink detection based on eye aspect ratio mapping (earm)," *Cognitive Robotics*, vol. 2, pp. 50–59, 2022.

[12] B. Mandal, L. Li, G. S. Wang, and J. Lin, "Towards detection of bus driver fatigue based on robust visual analysis of eye state," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 3, pp. 545–557, 2016.

[13] Y. Liu and X. Liu, "Spoof trace disentanglement for generic face anti-spoofing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3813–3830, 2022.

[14] F. M. Sukno, S.-K. Pavani, C. Butakoff, and A. F. Frangi, "Automatic assessment of eye blinking patterns through statistical shape models," in *International Conference on Computer Vision Systems*. Springer, 2009, pp. 33–42.

[15] D. Torricelli, M. Goffredo, S. Conforto, and M. Schmid, "An adaptive blink detector to initialize and update a view-basedremote eye gaze tracking system in a natural scenario," *Pattern Recognition Letters*, vol. 30, no. 12, pp. 1144–1150, 2009.

[16] M. Divjak and H. Bischof, "Eye blink based fatigue detection for prevention of computer vision syndrome." in *MVA*, 2009, pp. 350–353.

[17] Q. Wang, J. Yang, M. Ren, and Y. Zheng, "Driver fatigue detection: a survey," in *2006 6th world congress on intelligent control and automation*, vol. 2. IEEE, 2006, pp. 8587–8591.

[18] T. Drutarovsky and A. Fogelton, "Eye blink detection using variance of motion vectors," in *European conference on computer vision*. Springer, 2014, pp. 436–448.

[19] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing ai created fake videos by detecting eye blinking," in *2018 IEEE International workshop on information forensics and security (WIFS)*. IEEE, 2018, pp. 1–7.

[20] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 8261–8265.

[21] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2019, pp. 46–52.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[24] Y. Rao and J. Ni, "A deep learning approach to detection of splicing and

copy-move forgeries in images," in *2016 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 2016, pp. 1–6.

[25] N. Rahmouni, V. Nozick, J. Yamagishi, and I. Echizen, "Distinguishing computer graphics from natural images using convolution neural networks," in *2017 IEEE workshop on information forensics and security (WIFS)*. IEEE, 2017, pp. 1–6.

[26] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in *2018 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 2018, pp. 1–7.

[27] H. H. Nguyen, J. Yamagishi, and I. Echizen, "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)*. IEEE, 2019, pp. 2307–2311.

[28] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)*. IEEE, 2018, pp. 1–6.

[29] B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer," in *Proceedings of the 4th ACM workshop on information hiding and multimedia security*, 2016, pp. 5–10.

[30] D. Cozzolino, G. Poggi, and L. Verdoliva, "Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection," in *Proceedings of the 5th ACM workshop on information hiding and multimedia security*, 2017, pp. 159–164.

[31] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.