Deepfake detection using Deep Learning

Harika Golusu
Aditya Institute of Technology and Management, Tekkali , AP
golusuharika@gmail.com

Bejjipuram Naveen
Aditya Institute of Technology and Management, Tekkali , AP
naveenbejjipuram14@gmail.com

Akash Sanapala
Aditya Institute of Technology and Management, Tekkali , AP
sanapalaakash@gmail.com

Ganesh Neelapu
Aditya Institute of Technology and Management, Tekkali , AP
neelapuganesh752@gmail.com

Mrs. Tamada SriKanya
Assistant Professor , CSE-AIML
Aditya Institute of Technology and Management, Tekkali , AP
Srikanya659@gmail.com

Abstract - The rapid development of deepfake generation techniques raises severe concerns about digital media authenticity, privacy, and misinformation. This work proposes an effective deepfake detection framework in videos using deep learning techniques that fuse spatial and temporal feature learning. The proposed approach leverages a pre-trained ResNeXt convolutional neural network for robust spatial feature extraction from facial frames, followed by a Long Short-Term Memory (LSTM) network that captures temporal dependencies across video sequences. Facial regions first undergo preprocessing, which includes frame extraction, face detection, and face cropping, to ensure that the model focuses on relevant facial information only. Transfer learning is applied to make training more effective and perform well on limited datasets. A dataset consisting of real and fake videos is used for evaluating the model. The experiments are conducted by taking different numbers of frames per video. From the results, it is established that an increase in temporal information enhances detection accuracy considerably. A maximum of 93.58% accuracy is achieved for 100 frames per video. Moreover, the trained model is integrated into a Django-based web application where users can upload videos and get real-time predictions for deepfakes. Experimental results show that the proposed ResNeXt-LSTM framework is highly effective, scalable, and

Keywords - Deepfake Detection, Deep Learning, ResNeXt, Long Short-Term Memory (LSTM), Video Classification, Transfer Learning, Face Detection, Media Forensics.

I. INTRODUCTION

Due to the rapid development of deep learning, the creation of highly realistic synthetic media, also referred to as deepfakes is possible, and these synthetic media alter or generate video content by substituting or altering faces. Although these technologies can also be used positively to enhance entertainment and creation of digital contents, they are also very dangerous to privacy, security, and the trust of people. Misinformation, identity theft, political manipulation, and social engineering attacks, in which deepfakes have been used more and more extensively, are putting a strong need on a reliable detection system in current digital media forensics. Deepfake detection is especially difficult with video data because video data are time-based, and the quality of the manipulation methods is growing. In contrast to image-based deepfakes, video deepfakes use the consistency in time to create more realistic forged content. The old ways of forensic tools are no more that rely on handcrafted features to identify such advanced manipulations. Consequently, deep learning-based methods with the ability to learn discriminative spatial and temporal features automatically have received a lot of interest in the current research. This paper offers a deep learning architecture to detect deepfake videos using the integration of both convolutional neural networks and recurrent neural networks. The method employs a ResNeXt convolutional neural network which has been trained using the pre-trained features and is used as a feature extractor to extract fine-grained spatial features within facial regions in single frames of the video. Transfer learning is used to harness the representational strength of the already trained model, which saves on training time and enhances generalization. To graphically describe the time-related relationships between consecutive frames, a Long Short-Term Memory (LSTM) network will be employed, and it allows the system to learn motion-related irregularities and time effects that are prevalent in the manipulated videos. Before the model can be trained and inferred over, an additional preprocessing pipeline is run on the input videos. This pipeline involves dividing videos into frames, face detection, face cropping and the building of face-only video streams. The system is able to limit the amount of background noise and enhance the detection accuracy through the use of facial information only. The trained and tested deepfake detection model is then trained and tested on the processed data with varying temporal settings and the effects of the length of frame sequences on classification performance studied. The usefulness of the suggested framework is proved based on the vast range of experiments performed on the dataset of both real and faked videos. The findings show that, adding more frames can dramatically increase the accuracy of the detection, proving the usefulness of the temporal modeling in video-based deepfake detection. In addition, the trained model is incorporated in a Django-based web-application that enables the user to upload video and receive real-time predictions, which emphasizes the practicability of the system. Altogether, this study offers a scalable and effective deepfake detection framework that would consolidate a robust spatial feature response with a temporal sequence learning system, which would fit a real-world implementation scenario.

II. RELATED WORK

Deepfake content detection has been a dynamic field of research because of the rapid advancement of generative models including Generative Adversarial Networks (GANs) and autoencoders. The original methods to detect deepfakes used handcrafted features that detected visual artifacts, color irregularities, and frequency-domain abnormalities. Although useful with the early-generation deepfakes, these traditional forensic methods were not generalizable to high-quality manipulations and very sensitive to compression and other post-processing steps. CNNs have become common in detecting image-based deepfake with the development of deep learning. Some of them involved deep CNN architectures to identify spatial artifacts of altered facial images, which include blending and unnatural textures and facial mark inconsistencies. Preserving pre-trained models VGG, Inception, and Xception models have shown good performance after being fine-tuned on deep fake datasets. Nevertheless, the above approaches were mainly based on frame-level classification and did not resolve temporal errors between video sequences. In order to overcome the shortcomings of the frame-based solutions, researchers started to consider the methods of video-level deepfake detection that use the time modeling. Recurrent neural networks have been used to learn time dependencies among successive frames, especially Long Short-Term memory (LSTM) networks.

Through the combination of CNN-based spatial feature extraction with LSTM-based temporal analysis, these hybrid models have been shown to be more robust in detecting the subtle temporal features like unnatural eye-blink patterns, facial motion inconsistencies, and frame-to-frame distortions, which are put in place during video synthesis. Recent papers have considered the application of new CNN models like ResNet and ResNeXt to the problem of deepfake detection because of their powerful feature representation and training efficiently with residual connections. In particular, ResNeXt, offers the concept of cardinality, which enables the network to acquire a wide range of different feature representations without sacrificing computational efficiency. ResNeXt has also been applied together with sequence models, and studies have shown that ResNeXt can perform better than traditional CNN-LSTM pipelines in complex and high-resolution video inputs. Also, there have been a number of studies which have concentrated on preprocessing techniques aimed at improving the performance of the models by isolating the face regions in videos. Face detection and face cropping methods are used to minimise the amount of background noise, as well as to make sure that the model is focused on the most affected areas of the video. It has been demonstrated that such preprocessing pipelines substantially enhance the accuracy of detection and model generalization between dissimilar datasets.

Although these have been developed, most of the current methods have problems with scalability, real-time inferences, and implementation in real-world contexts. Moreover, some previous studies have not addressed the issue of performance of models with the temporal sequence length extensively. These gaps have prompted the proposed work to combine a ResNeXt-based feature extractor with an LSTM-based temporal classifier, to test the performance of the proposed system on different numbers of frames, and to apply the trained model to an actual web-based application framework, thus closing the gap between research and real-world application.

III. METHODOLOGY

In this section, the entire methodology of the deepfake video detection approach based on a deep learning-based framework that incorporates both the space and time features learning is explained. The suggested methodology consists of a pipeline with structured steps, involving pre-processing of data, feature extraction with the help of a convolutional neural network, temporal sequence prediction with the help of a recurrent neural network, and video-level classification. Video preprocessing will be the initial step in the methodology. The individual frames are extracted out of each input video so as to allow frame-wise analysis. Each frame is then subjected to face detection in order to identify and isolate the facial region, since facial regions are the most widely manipulated parts of the deepfake video. Identified faces are cropped and resized into a standard fixed resolution so that there is uniformity in the dataset. This step of preprocessing is used to minimize background noise and makes sure that the model is center of attention to facial features, which will enhance the detection accuracy and robustness. After preprocessing, spatial feature extraction is done using a pre-trained ResNeXt convolutional neural network. ResNeXt is used as a feature extractor, as opposed to training a CNN directly, and where the transfer learning is used, based on the rich representations that have been trained on large image datasets. Individual cropped face frames are run through the ResNeXt network and high-level feature vectors are obtained by deriving the ultimate convolutional layers. These feature vectors pick up the fine spatial artifacts like the texture disparities, blending mistakes and unnatural facial features which are indicative of tampered content. The extracted feature vectors are arranged as sequences and sent through a Long Short-Term Memory (LSTM) network in order to extract temporal dependence among video frames. The LSTM is ideally suited to the sequential representation and the study of long-range dependencies of time-dependent information. Through the study of the overall changes in facial features over a period, the LSTM can identify inconsistencies with time including abnormal facial movements and distortion of frames on frames of deepfake videos but not on real ones. The LSTM layer output is inputted into a fully connected classification layer, and a sigmoid activation to carry out binary classification. Depending on the spatial-temporal representations learnt, the model predicts an input video as real or fake. The model is trained with the help of a binary cross-entropy loss function, and the performance is measured with the help of accuracy, which is the main parameter. The experiment is done with varying number of frames per video to determine the impact of the temporal information on the detection performance. The capacity of the model to exploit time series is tested in a systematic manner by changing the sequence length. The trained model is then deployed to a Django-based web application, which allows users to submit videos and get the results of deepfake detection in real time. This end-to-end approach also guarantees high level of detection performance and workability in a real world context.

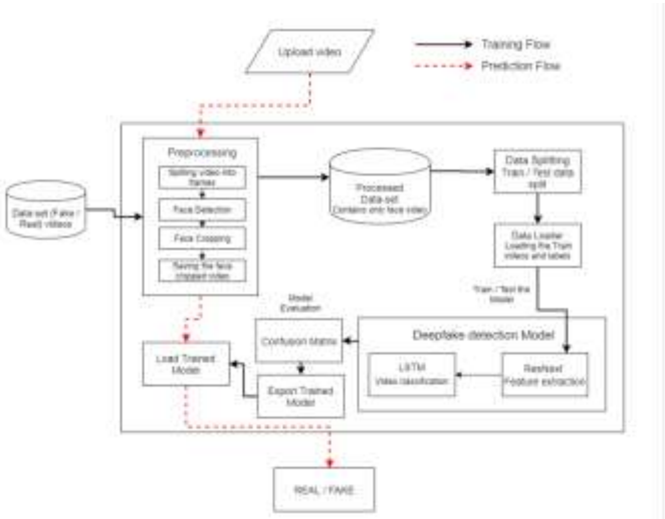

Upload video
Training Flow
Prediction Flow
Data set (Fake / Real) videos
Preprocessing
Splitting video into frames
Face Detection
Face Cropping
Saving the face cropped video
Processed Data-set (Contains only face video)
Data Splitting Train / Test data split
Data Loader Loading the Train videos and labels
Train / Test the Model
Model Evaluation
Confusion Matrix
Deepfake detection Model
LSTM Video classification
ResNext Feature extraction
Load Trained Model
Export Trained Model
REAL / FAKE

Fig1. System Architecture

The proposed deepfake detection framework architecture is an end-to-end pipeline, which takes in input video and extracts meaningful spatial and temporal features and delivers a correct classification output. The architecture brings together video preprocessing, deep learning-based feature extraction, temporal sequence modeling, and a web-based application to make real-time inference. The user interface layer starts with the user interface layer which is made up of a Django based web application. This layer enables end users to post a video file with the help of an easy and user-friendly interface. A video uploaded is sent to the backend processing module, and any other process occurs automatically, without any human intervention. At the video preprocessing module, the video uploaded is initially broken down into individual frames. In every frame, face detection is done to determine the facial area since the manipulations of deepface mainly focus on facial features. The identified faces are cut and rescaled to the desired constant size in order to have uniformity in all the inputs. This will eliminate background noise and pass only pertinent information on facial parts to the deep learning model. The already processed face frames are then inputted into a spatial feature extraction module which is powered by a pre-trained ResNeXt convolutional neural network. ResNeXt has been used as a feature extractor instead of doing direct classification, and it produces high-dimensional feature vectors representing each frame. The characteristics of these feature vectors are significant spatial features including texture patterns, face structure, and any visual artifact that can be a sign of manipulation. Transfer learning can be used to make the system use the powerful representations that it learned using large-scale datasets to enhance its efficiency and accuracy. The obtained feature vectors are then arranged in a sequential format and fed to the temporal modeling module that was done through a Long Short-Term Memory (LSTM) network. The LSTM examines the temporal dynamics of facial features between successive frames, enabling the system to detect both motion-based inconsistencies and temporal artifacts which are hard to detect through the analysis of single frames alone. Such a spatial and temporal learning combination increases the strength of the detection process markedly. The LSTM layer is connected to the classification module that is made up of fully connected layers and and a sigmoid activation function. This module generates a binary prediction that depicts whether the video input is real or deep fake. The end product of the prediction is then re-transmitted to the Django application. Lastly, the presentation result layer shows the user the result of the classification in real time. The user gets a direct idea of whether the uploaded video is considered as a legitimate one or it is manipulated. This scalable and modular architecture guarantees the efficient processing, the detection accuracy is high and the system is easily deployed to the real-world environment.

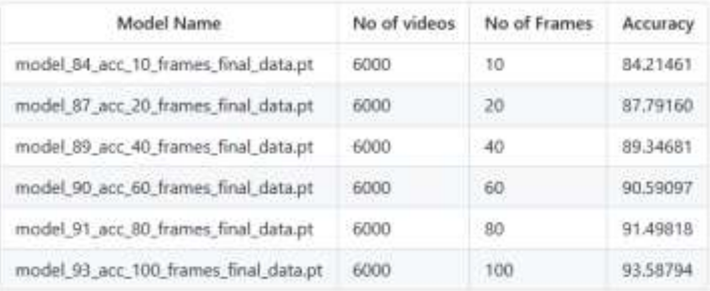
IV. EXPERIMENTAL RESULTS

This part provides the experimental analysis of the proposed deepfake detection model using the ResNeXt– LSTM architecture. The aim of the experiments will be both to evaluate the usefulness of spatial-temporal feature learning, and to measure the effects of the length of the temporal sequence on the performance of detection. Accuracy is considered the main performance measure to evaluate the model since the task is performed with binary classification of the real and deepfake videos.

The experiments are carried out by varying the quantity of frames obtained out of each video to investigate how the temporal information affects the accuracy of classification. In every configuration, facial frames are extracted, preprocessed, and subjected to the ResNeXt network to extract spatial features after which, temporal modeling is done using the LSTM layer. The same training and evaluation plan is kept throughout all experiments in order to compare them fairly.

The findings show that there is a definite increase of the detection of the performance with the increase of the number of frames per video. By using fewer frames, the model is mostly dependent on less temporal data, and thus it is unable to detect fine motion-based inconsistencies. The higher the rate of frames, the more the LSTM can learn the richer time pattern resulting into more accurate differentiation of real and manipulated videos. A maximum detection rate of 93.58 per cent is attained at 100 frames per video, thus demonstrating that longer temporal sequence dramatically improves the performance of a model. The results indicate the significance of time modeling in deepfake detection in videos. Though spatial characteristics learned using the ResNeXt network are used to successfully detect frame-level artifacts, the temporal analysis of the LSTM is important in detecting inconsistencies between frames that are typical of deep fake videos. These two elements combined leads to a strong and solid detection framework.

Besides the quantitative performance evaluation the trained model is tested in a real-world situation by integrating it into a web application based on Django. The system manages to take user uploaded videos and offer correct predictions in real time, which proves the practicality and scalability of the suggested approach. All in all, the experimental findings prove the fact that ResNeXtLSTM framework can be used to effectively detect deepfakes in videos and that adding longer temporal sequences will result in a substantial rise in the quality of classification.



Model Name	No of videos	No of Frames	Accuracy
model_84_acc_10_frames_final_data.pt	6000	10	84.21461
model_87_acc_20_frames_final_data.pt	6000	20	87.79160
model_89_acc_40_frames_final_data.pt	6000	40	89.34681
model_90_acc_60_frames_final_data.pt	6000	60	90.59097
model_91_acc_80_frames_final_data.pt	6000	80	91.49818
model_93_acc_100_frames_final_data.pt	6000	100	93.58794

Table1. Results

The results table presents the deepfake detection accuracy obtained using the proposed ResNeXt–LSTM framework for different numbers of frames extracted from each video. The table clearly demonstrates a progressive improvement in classification accuracy as the temporal sequence length increases. When a smaller number of frames is used, the model achieves comparatively lower accuracy due to limited temporal context, restricting the LSTM's ability to capture motion-based inconsistencies.

As the number of frames per video increases, the detection accuracy improves significantly. This improvement indicates that the LSTM network benefits from longer frame sequences, enabling it to learn more stable and discriminative temporal patterns associated with deepfake manipulations. The highest accuracy of **93.58%** is achieved when 100 frames per video are used, confirming that richer temporal information plays a crucial role in enhancing detection performance.

The results also validate the effectiveness of combining spatial feature extraction using ResNeXt with temporal modeling using LSTM. While ResNeXt efficiently captures frame-level artifacts, the LSTM complements it by identifying temporal inconsistencies across consecutive frames. Overall, the table highlights the strong correlation between temporal depth and model accuracy, reinforcing the importance of temporal sequence modeling in video-based deepfake detection.

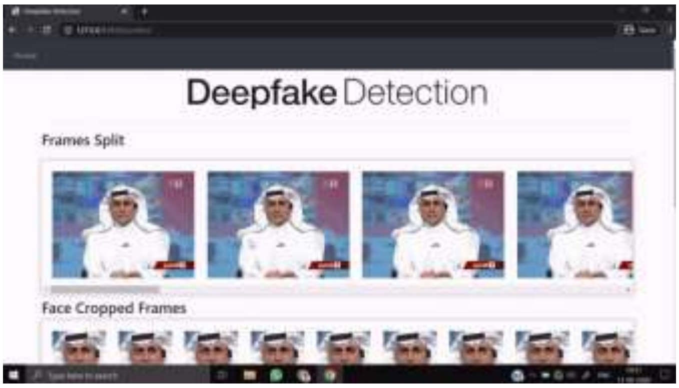


Fig 2. UI of the project

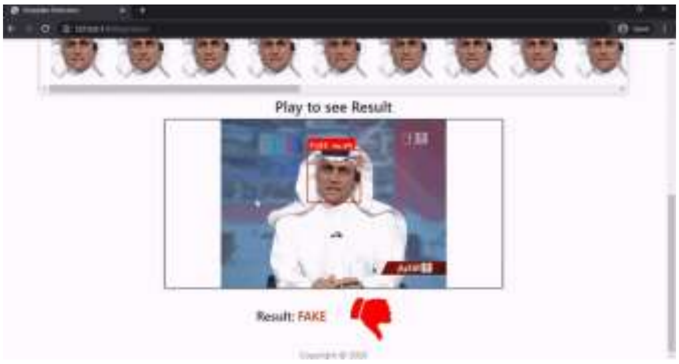


Fig 3.Result of the application

To demonstrate the practical applicability of the proposed deepfake detection framework, the trained ResNeXt–LSTM model is deployed using a Django-based web application. The web interface allows users to upload a video file, which is then processed by the backend deep learning pipeline to generate a real-time prediction.

As shown in the interface, the uploaded video is first decomposed into individual frames, which are displayed under the *Frames Split* section. This visualization confirms the correct extraction of frames from the input video. Subsequently, face detection and face cropping are applied to each frame, and the extracted facial regions are displayed under the *Face Cropped Frames* section. This step verifies that the preprocessing pipeline accurately isolates facial information, which is critical for effective deepfake detection.

Once preprocessing is completed, the extracted facial frames are passed through the trained ResNeXt network for spatial feature extraction, followed by temporal analysis using the LSTM model. During inference, the system highlights the detected face region within the video frame using a bounding box and overlays the predicted label. In the displayed result, the model successfully classifies the input video as **FAKE**, with a high confidence score, clearly indicating the presence of manipulated content.

The final classification result is prominently displayed below the video player, providing a clear and user-friendly output. This deployment confirms that the proposed deep learning model is not only effective in offline evaluation but also capable of performing reliable deepfake detection in a real-world, interactive environment. The integration of preprocessing visualization and prediction output enhances transparency and interpretability, making the system suitable for practical use cases such as media verification and digital forensics.

V. Conclusion

The This paper has presented and developed a strong deepfake video detection system on the basis of deep learning in which a ResNeXt and a Long Short Term Memory (LSTM) network have been combined. It is a system that is effective in combining both spatial feature extraction and time sequence modeling to deal with the difficulties presented by the detection of advanced video-based deepfakes. The proposed method reduces the number of irrelevant information because it narrows down the regions of the face by analyzing the face using effective preprocessing pipeline, thereby increasing the chances of detection. The results of the experiment prove that ResNeXt as a feature extractor with LSTM as a temporal analyzer shows much better performance in detecting deepfakes. The analysis has revealed a strong association between the length of temporal sequence and classification accuracy and the model has the highest accuracy of 93.58% with longer frame sequences. These results demonstrate the significance of time modeling in detecting the subtle inconsistencies that cannot be readily identified using frame level analysis. In addition, the practical relevance of the suggested framework is proved by the fact that the trained model is applied to a web application using Django. The system has been successful in processing user-uploaded videos, visualizing prior processing activities and making correct real-time predictions which make it applicable in the real world context to verify media and digital forensics. In general, the present research confirms that advanced convolutional neural networks combined with recurrent neural networks can be successfully and effectively used to detect deepfakes in videos. The proposed framework is effective in both that it has a high level of detection as well as bridging the academic research and the actual world implementation which will play a role in curbing the ever increasing threats of deepfake technology.

VI. FUTURE WORK

Even though the suggested ResNeXt-LSTM-based deepfake detector framework shows a high level of performance and operational feasibility, there are multiple areas to be considered in order to improve the system. The possibility of inclusion of attention mechanisms in the temporal modeling phase is one of the possible improvements. Attention-based models can assist the system to concentrate on the most informative frames of a video sequence to be reduced and enhance the accuracy of the detection particularly in longer videos. It is also possible to investigate the application of transformers-based architecture to model temporal features in the future. Transformers have demonstrated better results in sequence learning problems and can potentially learn long-term temporal patterns better than the more classical recurrent networks. This could also be enhanced by adding vision transformers or hybrid CNN transformer models to enhance robustness against advanced methods of deepfake generation. The other crucial extension is a training and testing the model on bigger and more varied datasets of deepfakes produced via various techniques, resolutions and compression rates. This would make the system better capable of generalization and resistant to manipulation techniques that are harder to detect. Also, cross-dataset testing may be conducted in order to determine the practicality of the model. In terms of deployment, there are future improvements that can be made to the model in order to provide real-time processing on edge devices using model compression algorithms including pruning and quantization. This would allow it to be deployed in resource constrained systems like mobile devices or surveillance systems. Lastly, the system may be expanded to multimodal deep fake detection including audio analysis and lip-sync consistency check. Visual and audio signals could be used to create a more efficient and efficient deepfake detection paradigm, which could be used even more to reinforce the importance of digital media forensics and security-sensitive tasks.

VI. REFERENCES

[1] I. Goodfellow et al., "Generative Adversarial Nets," in Proc. Advances in Neural Information Processing Systems (NeurIPS), Montreal, Canada, 2014, pp. 2672–2680.

[2] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A Compact Facial Video Forgery Detection Network," in Proc. IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, 2018, pp. 1–7.

[3] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 1251–1258.

[4] A. Rossler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," in Proc. IEEE International Conference on Computer Vision (ICCV), Seoul, South Korea, 2019, pp. 1–11.

[5] J. He, X. Liu, and S. Jain, "Deepfake Detection Using Recurrent Neural Networks," IEEE Access, vol. 8, pp. 121–130, 2020.

[6] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 1492–1500.

[8] Y. Li, M.-C. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking," in Proc. IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, 2018.

[9] H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task Learning for Detecting and Segmenting Manipulated Facial Images and Videos," in Proc. IEEE International Conference on Biometrics (ICB), Gold Coast, Australia, 2019.

[10] Y. Li and S. Lyu, "Exposing DeepFake Videos By Detecting Face Warping Artifacts," in Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 2019.

[11] A. Sabir et al., "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos," in Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019.

[12] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in Proc. International Conference on Learning Representations (ICLR), 2015.

[13] Y. Guera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," in Proc. IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS), Lecce, Italy, 2018, pp. 1–6.

[14] T. M. Mitchell, Machine Learning. New York, NY, USA: McGraw-Hill, 1997.

[15] J. Donahue et al., "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Communications of the ACM, vol. 60, no. 6, pp. 84–90, 2017.

[17] D. Cozzolino, J. Thies, A. Rössler, C. Riess, M. Nießner, and L. Verdoliva, "ForensicTransfer: Weakly-Supervised Domain Adaptation for Forgery Detection," IEEE Transactions on Information Forensics and Security, vol. 14, no. 6, pp. 1524–1536, 2019.

[18] Z. Tolosana et al., "DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection," Information Fusion, vol. 64, pp. 131–148, 2020.

[19] P. Korshunov and S. Marcel, "Deepfakes: A New Threat to Face Recognition? Assessment and Detection," arXiv preprint arXiv:1812.08685, 2018.

[20] A. Dosovitskiy et al., "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," in Proc. International Conference on Learning Representations (ICLR), 2021.