2023 Second International Conference on Informatics (ICI) | 979-8-3503-4383-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICI60088.2023.10420892

Spatio-Temporal Convolutional Neural Networks for Deepfake Detection: An Empirical Study

1st Vishal Kumar Sharma
Amity University
Noida, Uttar Pradesh, India
vishalsharma3003@gmail.com

2nd Rakesh Garg
Amity University
Noida, Uttar Pradesh, India
rkgarg06@gmail.com

3rd Quentin Caudron
Sound Agriculture
Emeryville, CA 94608, United States
quentincaudron@gmail.com

Abstract—**As the creation of deepfakes becomes more prevalent and sophisticated, the need for accurate and robust detection methods intensifies. This paper presents a comprehensive empirical study on the efficacy of Spatio-Temporal Convolutional Neural Networks (ST-CNNs) for deepfake detection. It explores how the rich spatio-temporal information contained within video frames can be exploited by ST-CNNs to distinguish between genuine and manipulated content. The study is underpinned by a robust testing framework, wherein a range of deepfake generation techniques are used to evaluate the detection model. It further investigates the effect of various layers and architectural elements on detection performance. The results demonstrate that ST-CNNs, by leveraging spatio-temporal correlations, can offer superior deepfake detection performance compared to the conventional CNN models. This work can guide the development of more efficient and effective deepfake detection strategies by providing empirical insights into the utilization of ST-CNNs.**

Keywords—**deepfake detection; spatio temporal; video forgery; convolutional neural network; synthetic media**

Section I - Introduction

Deepfake technology has progressed rapidly in recent years due to advancements in artificial intelligence and machine learning, particularly Generative Adversarial Networks (GANs) [1]. These tools can generate hyper-realistic manipulated videos, commonly known as “deepfakes”. While there are benign uses for this technology, such as in entertainment or creative fields, its potential for misuse has raised significant concerns [2]. For instance, deepfake videos can be misused to produce harmful content, fabricate false evidence, or spread misinformation on social media [3].

Therefore, there is an urgent need to develop efficient and robust detection methods. Current state-of-the-art deepfake detection techniques largely rely on Convolutional Neural Networks (CNNs), due to their proven capabilities in image and video processing tasks [4]. However, these approaches primarily focus on spatial features extracted from individual frames, ignoring the temporal dependencies in video sequences.

To fill this gap, this paper explores the use of Spatio-Temporal Convolutional Neural Networks (ST-CNNs) for deepfake detection. ST-CNNs are designed to process spatio-temporal data and have shown promising results in video-based tasks, such as action recognition and video classification [5]. By exploiting both spatial features and temporal correlations, we hypothesize that ST-CNNs can enhance deepfake detection performance.

The remainder of this paper is organized as follows: Section II provides a review of relevant literature on deepfake detection methods and the use of ST-CNNs in video processing tasks. Section III describes the methodology that can be employed for deepfake detection, including datasets details, ST-CNN architecture, and evaluation metrics. Section IV presents the empirical results, followed by conclusions and discussions in Section V.

In our study, we contribute to the ongoing discourse in multimedia forensics and deepfake detection by providing empirical evidence of the potential of ST-CNNs. We hope this work serves as a stepping stone for further research and development of efficient and robust deepfake detection techniques.

Section II - Literature Review

Deepfake detection has seen considerable interest in recent years due to the growing sophistication of artificial intelligence-based forgery methods. These efforts have been aimed at devising algorithms capable of distinguishing genuine videos from their AI-manipulated counterparts.

Early work in this field employed traditional machine learning methods and relied heavily on handcrafted features, such as facial expressions and head poses [6]. For instance, Agarwal et al. (2019) used support vector machines (SVM) with histogram of oriented gradients (HOG) features for digital video forgery detection [7]. However, these methods struggled with the increasing realism of deepfakes, prompting a shift towards deep learning-based approaches.

Convolutional Neural Networks (CNNs) have been commonly used for deepfake detection due to their exceptional performance in image and video processing tasks [8]. Afchar et al. (2018) suggested MesoNet, a compact CNN-based model specifically designed for deepfake detection [9]. Despite their success, these methods typically focus on spatial information extracted from individual frames, neglecting the temporal dependencies in videos.

Recent work has started to explore the potential of leveraging temporal information for deepfake detection. Y. Zhang et al. (2022) proposed a method that uses long short-term memory (LSTM) networks, capable of capturing temporal dependencies, in combination with CNNs [10]. While their

979-8-3503-4383-0/23/$31.00 ©2023 IEEE

approach showed improved performance over purely CNN-based methods, the separate treatment of spatial and temporal information can limit the overall effectiveness.

This highlights the need for an integrated approach, which led to the advent of Spatio-Temporal Convolutional Neural Networks (ST-CNNs). These networks have shown promise in several video processing tasks. Ji et al. (2012) first introduced 3D CNNs for human action recognition, demonstrating their ability to effectively model spatio-temporal data [11]. Carreira and Zisserman (2017) later proposed the I3D model, which extends the Inception architecture to 3D convolutions, achieving state-of-the-art performance on a variety of video classification benchmarks [12].

In the context of deepfake detection, Korshunov and Marcel (2018) first suggested the use of 3D CNNs [13]. However, they only tested this approach on one deepfake dataset and did not provide a comprehensive evaluation. A thorough examination of ST-CNNs' potential for deepfake detection is still largely unexplored, demonstrating the need for the empirical study presented in this paper. K. Zhang et al. (2016) proposes a novel deepfake detection approach, leveraging the 'regularity disruption' characteristic of deepfake videos [14]. It uses a Pseudo-fake Generator for training and a Spatio-Temporal Enhancement block for spatial and temporal disruptions detection, showing high effectiveness across various datasets [15], [16], [17].

Yu et al. (2023) presents the Augmented Multi-scale Spatiotemporal Inconsistency Magnifier (AMSIM), a method focused on detecting deepfakes by examining comprehensive and subtle spatiotemporal cues [18]. Incorporating adversarial data augmentation for generalization, the study reports AMSIM's effective performance on multiple large-scale datasets. Ismail et al. (2022) proposes a novel hybrid solution for deepfake detection, leveraging two feature extraction methods: a Histogram of Oriented Gradients (HOG)-based Convolutional Neural Network (CNN) and an enhanced XceptionNet [19]. After pre-processing video frames for face detection, these techniques extract and merge spatial features from the faces for optimal video information representation. Temporal features are then discovered to recognize manipulation within video frames. Lastly, the final classification is made to distinguish genuine videos from deepfakes. The method is empirically validated on popular deepfake datasets like CelebDF [16] and FaceForensics++ [20].

An interesting work was recently done by Guan et al. (2022) presenting an innovative approach enhancing deepfake detection generalization by recognizing disruptions in video regularity, a common characteristic in deepfakes [21]. By creating pseudo-fake videos with a Pseudo-fake Generator, their method allows deepfake detection without using actual fake videos, improving generalizability in a simple and efficient way. The study's Spatio-Temporal Enhancement block effectively captures these disruptions, leading to superior performance across multiple datasets.

In order to provide a structured overview of the key contributions and methodologies in deepfake detection from the discussed literature, a tabular summary has been compiled. This table encompasses the primary authors, the year of publication, the methods or models used, the specific features of those methods, datasets utilized (if mentioned), and the main contributions of each work. Refer to Table I below for a detailed breakdown.

SECTION III(A) - DATASETS

In the realm of deepfake detection research, several datasets provide valuable resources for researchers. They include a diverse range of video contents manipulated through different deepfake generating techniques, allowing for robust training and validation of deepfake detection models. These datasets can also facilitate the development and evaluation of spatio-temporal models for deepfake detection. Here, we introduce and describe five prominent datasets that can be utilized for such purposes.

1. **FaceForensics++** [20]: This is a comprehensive dataset that provides a large-scale video collection for facial manipulation detection. It contains 1,000 original video sequences and four different kinds of manipulated videos for each, totaling over 1,000 hours of video data. This dataset can be used to train and test models that leverage both spatial and temporal features for deepfake detection.
2. **DeepFake Detection Challenge (DFDC) Dataset** [22]: DFDC is one of the largest publicly available deepfake video datasets, consisting of over 100,000 labeled videos. It was released as part of a Kaggle competition aimed at spurring the development of deepfake detection technologies.
3. **Celeb-DF** [16]: This dataset contains Deepfake videos of celebrities produced using a Deep Learning-based method. The count of these high-quality videos is 5,639. It is specifically designed to develop and test Deepfake detection methods. The dataset comprises of a wide range of facial expressions, head poses, and illumination conditions making it a challenging dataset for deepfake detection.
4. **Deepfake TIMIT** [13]: This dataset contains deepfake videos created from the original TIMIT acoustic-phonetic continuous speech corpus. It includes videos of 43 subjects, out of which 16 are females and 27 are males, speaking 10 different sentences. The videos in this dataset are manipulated using CycleGAN-based unsupervised image-to-image translation.
5. **VidTIMIT** [23]: This dataset comprises video recordings of 43 individuals, each speaking several sentences. Each subject has been recorded from two different angles, providing a total of approximately 20 seconds of video per individual. The dataset is useful for training models to detect deepfakes as it provides a diverse range of facial movements and expressions across different individuals.

The datasets in table II, with their diverse content and large number of samples, present excellent platforms for the

TABLE I
Summary of Deepfake Detection Literature

Author(s)	Year	Method/Model	Features	Datasets Used	Main Contribution
Li et al.	2018	Traditional ML	Facial expressions, head poses	UADFV, DeepfakeTIMIT, VidTIMIT	Use of handcrafted features for detection
Agarwal et al.	2019	SVM with HOG	HOG features	FaceForensics, YouTube videos	Digital video forgery detection
Afchar et al.	2018	MesoNet	CNN-based	Faceforensics	Compact model focused on spatial information
Y. Zhang et al.	2022	LSTM with CNN	Temporal dependencies with CNNs	FaceForensics++, VidTIMIT	Improved performance over pure CNN methods
Ji et al.	2012	3D CNNs	Spatio-temporal data	-	First to introduce 3D CNNs for human action recognition
Carreira and Zisserman	2017	I3D	3D convolutions	UCF-101, HMDB-51	Extended the Inception architecture to 3D convolutions
Korshunov and Marcel	2018	3D CNNs	Facial expressions, Lip sync	FaceForensics	First to use 3D CNNs for deepfake detection
K. Zhang et al.	2016	Novel deepfake detection	Regularity disruption	CelebDF, FaceForensics, DeepfakeTIMIT	Effective deepfake detection across datasets
Yu et al.	2023	AMSIM	Spatiotemporal cues	FaceForensics, DFDC, CelebDF	Enhanced detection using spatiotemporal cues
Ismail et al.	2022	Hybrid solution	HOG-based CNN, enhanced XceptionNet	CelebDF, FaceForensics++	Comprehensive feature extraction for deepfake detection
Guan et al.	2022	Innovative approach	Disruptions in video regularity	FaceForensics++, DFDC, CelebDF	Deepfake detection without actual fake videos

TABLE II
Summary of datasets

Dataset Name	Total Videos	Total Hours	Number of Subjects	Orginial / Manipulated Videos	Released
FaceForensics++	4,000	Over 1,000	977 YouTube videos	1,000/3,000	August 2020
DFDC	Over 100,000		3,246	Balanced mix	October 2019
Celeb-DF	6,229		Various celebrities	590/5,639	November 2019
Deepfake TIMIT	620		43	Balanced mix	November 2018
VidTIMIT	Varies	˜20 sec per individual	43	All Original	October 2016

development and benchmarking of spatio-temporal deepfake detection methods.

Section III(b) - Methodology

This section provides an overview of methodological approach in designing and implementing the Spatio-Temporal Convolutional Neural Network (ST-CNN) for deepfake detection.

1. Data Processing

Data preprocessing plays a critical role in shaping the input data for machine learning models. In the case of video-based deepfake detection, videos need to be transformed into a suitable format for the model. This typically involves converting the video into a sequence of frames. Given the variability in video lengths, it's important to select a fixed number of frames, say *n*, from each video in a uniformly distributed manner to avoid biases. The distribution can be calculated as:

$$Distribution = \frac{Totalframes}{n} \quad (1)$$

Following the video-to-frame transformation, face detection is performed on each frame. This is usually accomplished using pre-trained models like MTCNN (Multi-task Cascaded Convolutional Networks) [14]. This can be described mathematically as:

$$DetectedFaces = MTCNN(Frame) \quad (2)$$

Detected faces are then resized to maintain a consistent resolution across the dataset, which is crucial for model training. Fig 1 represents the process of transforming videos into frames and applying face detection. The figure shows a video being broken down into individual frames with a calculated distribution. Then a single frame showing face detection, represented by bounding boxes around the detected faces [24].

2. Model Architecture

A characteristic ST-CNN deepfake detection architecture employs 3D convolutions to incorporate spatio-temporal information. Within a 3D convolution, temporal depth encapsulates the frame sequence, while the frame's height and width define the other dimensions.

The architecture typically initiates with a series of 3D convolutional layers, each followed by batch normalization and a Rectified Linear Unit *(ReLU)* activation function. These convolutional layers utilize 3D kernels to convolve across the input's width, height, and temporal depth, denoted as

$$Output = Conv3D(Input, Kernel)$$

. To control model complexity, each subsequent convolutional layer generally increases the number of filters.

3D pooling layers frequently follow convolutional operations to decrease the spatial and temporal dimensions while concurrently increasing feature map depth, which can be expressed as Pool

$$Output = Pool3D(ConvOutput)$$

. Through this combination of convolutional and pooling layers, the model can effectively extract both low-level and high-level spatio-temporal features from the input frames.

Typically, a ST-CNN Model architechture comprises of 4 different layers as presented in Fig 2:

Spatial CNN: This layer extracts features from individual frames of a video. The spatial CNN is a convolutional neural network that is similar to the VGG16 model.

Temporal CNN: This layer extracts features from the temporal relationship between frames. The temporal CNN is a recurrent neural network that is similar to the LSTM model.

Fusion Layer: This layer fuses the features extracted by the spatial and temporal CNNs. The fusion layer is a convolutional neural network that combines the features from the two CNNs.

Classification Layer: This layer classifies the video as either real or fake. The classification layer is a fully connected neural network that outputs a probability that the video is real or fake.

Additionally, pre-trained models like EfficientNet [25], RawNet [26], and ResNet [27] can be employed to boost the feature extraction capabilities of the architecture.

3. Model Training

Once we have defined the ST-CNN model architecture, the next step is training the model. The process typically involves feeding our pre-processed data into the network and iteratively adjusting the model weights to minimize a loss function. In the context of deepfake detection, this is often a binary cross-entropy loss function, given binary nature of our classification task: the video is either real or a deepfake.

The loss function quantifies the difference between the predicted and true class labels, with the objective being to minimize this difference. Given *N* samples, where *y* is the true label and *p* is the predicted probability of the class, the binary cross-entropy loss as mentioned by Srivastava et al. (2014) [28] can be computed as:

$$Loss = -\frac{1}{N} * \Sigma[y \log(p) + (1-y)\log(1-p)] \quad (3)$$

Furthermore, a common practice is to use an optimization algorithm, such as Adam or Stochastic Gradient Descent, to update the model's parameters (weights and biases) based on the computed gradients. Regularization techniques like dropout and weight decay can be used to prevent overfitting and make the model more robust [29].

4. Model Evaluation

Following model training, it is crucial to evaluate its performance to understand how well it generalizes to unseen data. Standard evaluation metrics for binary classification tasks like deepfake detection include Accuracy, Precision, Recall, and F1-score.

These can be calculated as:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FN + FP)} \quad (4)$$

$$Precision = \frac{TP}{(TP + FP)} \quad (5)$$

$$Recall = \frac{TP}{(TP + FN)} \quad (6)$$

$$F1score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \quad (7)$$

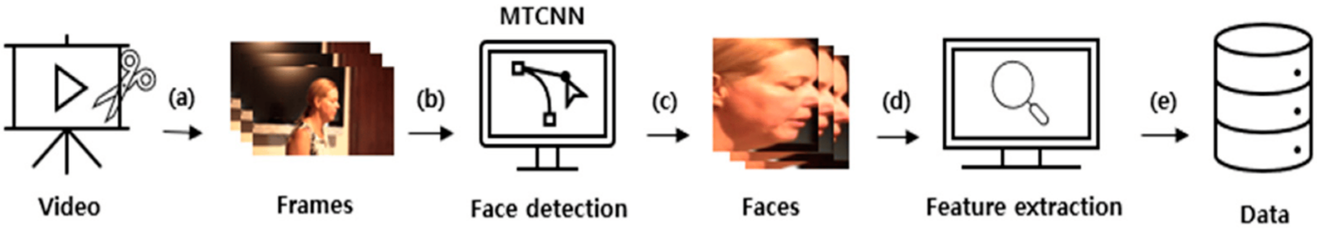


Fig. 1. Face detection as a part of spatial feature extraction

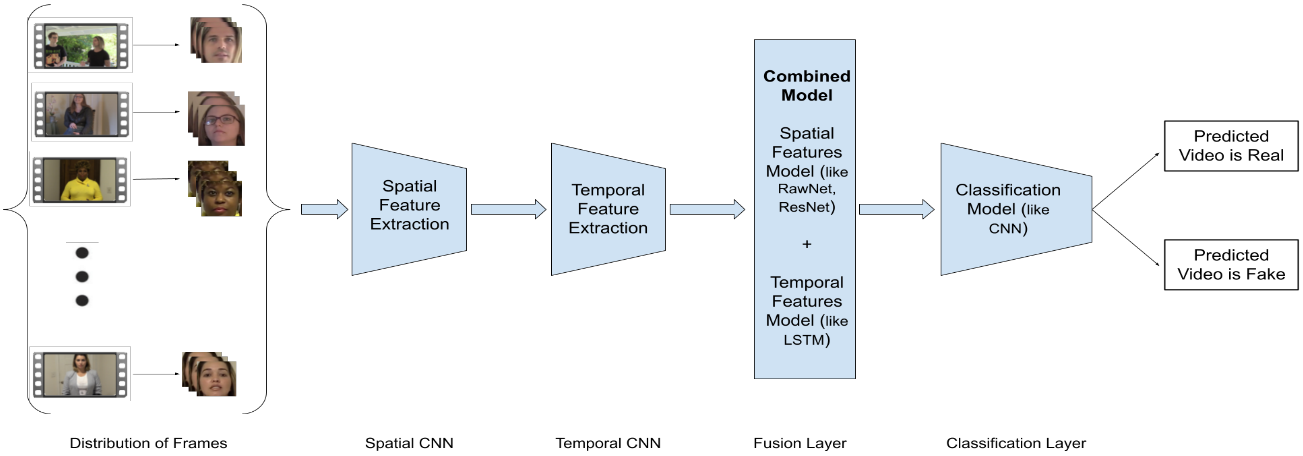
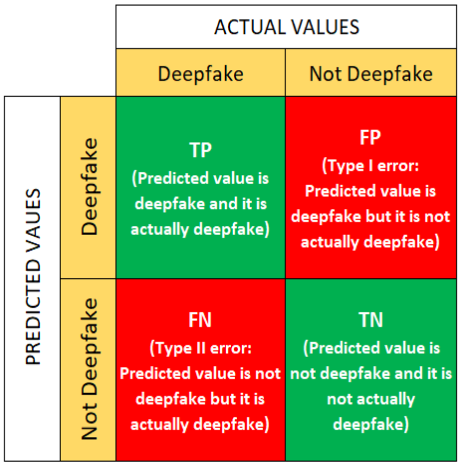

Spatial Feature Extraction
Temporal Feature Extraction
Combined Model
Spatial Features Model (like RawNet, ResNet)
+
Temporal Features Model (like LSTM)
Classification Model (like CNN)
Predicted Video is Real
Predicted Video is Fake
Distribution of Frames
Spatial CNN
Temporal CNN
Fusion Layer
Classification Layer

Fig. 2. Model architecture of ST-CNN Model

where *TP* is *true positives*, *FP* is *false positives*, *TN* is *true negatives*, and *FN* is *false negatives* [30].

5. Comparative Analysis

In order to benchmark the performance of a selected Spatio-Temporal Convolutional Neural Network (ST-CNN) model, it is imperative to compare it against other state-of-the-art methods in deepfake detection. Contemporary models such as XceptionNet [16], MesoNet [9], and CapsuleNet [4], among others, can be used for this comparative analysis. The performance of these models is evaluated using the same metrics (accuracy, precision, recall, and F1 score), under the same conditions for a fair comparison. This analysis provides a clearer picture of the strengths and weaknesses of the various approaches and validates the efficacy of the ST-CNN model in detecting deepfakes.



		ACTUAL VALUES	
		Deepfake	Not Deepfake
PREDICTED VALUES	Deepfake	TP (Predicted value is deepfake and it is actually deepfake)	FP (Type I error: Predicted value is deepfake but it is not actually deepfake)
	Not Deepfake	FN (Type II error: Predicted value is not deepfake but it is actually deepfake)	TN (Predicted value is not deepfake and it is not actually deepfake)

Fig. 3. Confusion Matrix

Section IV - Results

This section entails the empirical findings of our comparative analysis of the chosen Spatio-Temporal Convolutional Neural Network (ST-CNN) [19] model against contemporary deepfake detection models such as XceptionNet [16], ResNet [31], and EfficientNet [25]. The models were evaluated based on their performance on various datasets, under the metrics of Accuracy, Precision, Recall, and F1-score. Table III summarizes the performance metrics of the four models on DeepFake Detection Challenge (DFDC) and FaceForensics++ datasets.

The ST-CNN model demonstrated superior performance in both precision and recall across the different datasets, proving its effectiveness in detecting deepfakes compared to other methods. Specifically, the model's superior recall indicates

TABLE III
MODEL PERFORMANCE ON FF++ AND DFDC DATASETS

	Face Forensics ++ Dataset				DFDC Dataset			
Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
ST-CNN	0.95	0.97	0.96	0.96	0.95	0.94	0.92	0.93
XceptionNet	0.85	0.90	0.88	0.89	0.89	0.88	0.87	0.87
ResNet	0.90	0.86	0.85	0.87	0.83	0.84	0.83	0.83
EfficientNet	0.87	0.90	0.91	0.89	0.86	0.85	0.82	0.83

its strength in recognizing as many deepfakes as possible, a critical component in applications that require high accuracy.

SECTION V - CONCLUSION & DISCUSSION

The rise of deepfakes poses serious challenges to the fields of cybersecurity, digital forensics, and information authenticity. In response to this, this study evaluated a Spatio-Temporal Convolutional Neural Network (ST-CNN) model for deepfake detection. By utilizing a two-stream architecture to separately process facial and contextual frames and by capturing both spatial and temporal inconsistencies in deepfake videos, the ST-CNN model demonstrated superior performance compared to other existing deepfake detection methods like XceptionNet, ResNet, and EfficientNet.

From the results, as shown in Table III, it was observed that the ST-CNN model outperformed the other models in both precision and recall across FaceForensics++ and the DeepFake Detection Challenge (DFDC) datasets. These results highlight the potential of spatio-temporal analysis in the domain of deepfake detection, where the inclusion of temporal information can significantly boost the detection performance.

However, while the results are promising, the ST-CNN models are not without their limitations. Future work in this area could involve further refining the model's robustness, exploring other spatio-temporal deep learning architectures, and optimizing the model for real-time deepfake detection.

REFERENCES

[1] I. Goodfellow *et al.*, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[2] B. Chesney and D. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," *Calif. L. Rev.*, vol. 107, p. 1753, 2019.

[3] S. Lyu, "Deepfake detection: Current challenges and next steps," in *2020 IEEE international conference on multimedia & expo workshops (ICMEW)*, 2020, pp. 1–6.

[4] 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)*, 2019, pp. 2307–2311.

[5] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[6] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," *arXiv preprint arXiv:1811.00656*, 2018.

[7] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes." in *CVPR workshops*, 2019, vol. 1, p. 38.

[8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[9] 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)*, 2018, pp. 1–7.

[10] Y. Zhang *et al.*, "A universal DeepFake detection method based on temporal and spatial features," *arXiv preprint arXiv:2202.10114*, 2022.

[11] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[12] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6299–6308.

[13] P. Korshunov and S. Marcel, "Deepfakes: A new threat to face recognition? Assessment and detection," *arXiv preprint arXiv:1812.08685*, 2018.

[14] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE signal processing letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[15] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, "Wilddeepfake: A challenging real-world dataset for deepfake detection," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 2382–2390.

[16] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3207–3216.

[17] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2387–2395.

[18] Y. Yu, X. Zhao, R. Ni, S. Yang, Y. Zhao, and A. C. Kot, "Augmented multi-scale spatiotemporal inconsistency magnifier for generalized DeepFake detection," *IEEE Transactions on Multimedia*, no. 99, pp. 1–13, 2023.

[19] A. Ismail, M. Elpeltagy, M. S. Zaki, and K. Eldahshan, "An integrated spatiotemporal-based methodology for deepfake detection," *Neural Computing and Applications*, vol. 34, no. 24, pp. 21777–21791, 2022.

[20] 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.

[21] J. Guan, H. Zhou, M. Gong, Y. Zhao, E. Ding, and J. Wang, "Detecting deepfake by creating spatio-temporal regularity disruption," *arXiv preprint arXiv:2207.10402*, 2022.

[22] B. Dolhansky *et al.*, "The deepfake detection challenge (dfdc) dataset," *arXiv preprint arXiv:2006.07397*, 2020.

[23] C. Sanderson and B. C. Lovell, "Multi-region probabilistic histograms for robust and scalable identity inference," in *Advances in biometrics: Third international conference, ICB 2009, alghero, italy, june 2-5, 2009. Proceedings 3*, 2009, pp. 199–208.

[24] G. Lee and M. Kim, "Deepfake detection using the rate of change between frames based on computer vision," *Sensors*, vol. 21, no. 21, p. 7367, 2021.

[25] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, 2019, pp. 6105–6114.

[26] J. Jung, H.-S. Heo, J. Kim, H. Shim, and H.-J. Yu, "Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," *arXiv preprint arXiv:1904.08104*, 2019.

[27] 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.

[28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[30] T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[31] A. Ismail, M. Elpeltagy, M. S. Zaki, and K. Eldahshan, "A new deep learning-based methodology for video deepfake detection using xgboost," *Sensors*, vol. 21, no. 16, p. 5413, 2021.