TITLE

B. Tech. Project Mid Sem Report

Submitted by

Sumit Sunil Girnar **112003045**

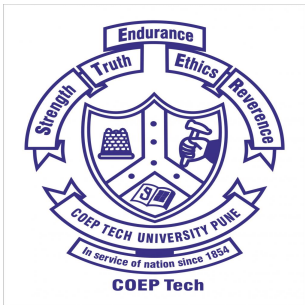
Swapnil Santosh Gite **112003046**

Om Prakash Gurav **112003047**

Under the guidance of

Prof. P. R. Deshmukh

COEP Technological University, Pune



DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

COEP TECHNOLOGICAL UNIVERSITY, PUNE-5

March 2024

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING,

COEP TECHNOLOGICAL UNIVERSITY, PUNE-5

CERTIFICATE

Certified that this project titled, "Deepfake Detection using PPG Signals" has been successfully completed by

Sumit Sunil Girnar	**112003045**
Swapnil Santosh Gite	**112003046**
Om Prakash Gurav	**112003047**

and is approved for the partial fulfillment of the requirements for the degree of "B.Tech. Computer Engineering".

SIGNATURE
NAME OF GUIDE
Project Guide
Department of CSE
COEP Tech Pune,
Shivajinagar, Pune - 5.

SIGNATURE
NAME OF HOD
Head
Department of CSE
COEP Tech Pune,
Shivajinagar, Pune - 5.

Abstract

The advancement of generative techniques for producing fake portrait videos poses a significant societal challenge, particularly with the emergence of realistic deep fakes used for political propaganda, imitating celebrities, and manipulating identities. This project aims to detect forged faces as a way to defend against the misuse of deepfakes and classify the videos as real or fake. By looking at subtle changes in skin colour caused by blood flow in the face using remote visual photoplethysmography (rPPG), past research suggests that normal heartbeat rhythms seen in real-life videos will be disturbed or completely broken in DeepFake videos. Taking advantage of these disruptions in facial colour changes, the rPPG signal becomes a strong biological indicator for effective deepfake detection. The proposed approach involves utilising a spatial-temporal PPG map to extract heartbeat signals from various facial regions and also enhancing the model through the incorporation of neural network techniques such as attention modules and transformers. Additionally, the study checks the model's strength by testing it on different datasets and looks into the impact of video compression on rPPG information from both real and fake videos.

Chapter 1

Introduction

The rise in advanced and accesible technology has led to a increase in deepfake videos across social media platforms, hence a significant challenge in modern society. Deepfakes involve digitally manipulating images or videos to replace a person's identity with another's which leds further spreading misleading information or creating fake news. Notably, deepfake technology has been misused to produce false videos of prominent public figures like Barack Obama and Joe Biden, worsening concerns about the spread of misinformation.. Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) have emerged as popular tools for creating deepfakes [2], leveraging large datasets to create realistic images and videos. Deep learning methods such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN),and Long Short-Term Memory (LSTM) have been proposed to address this challenge .There are different types of face manipulation techniques such as two computer graphics-based approaches (Face2Face and FaceSwap) and two learning-based approaches (DeepFakes and NeuralTextures).

Face2Face enables the transfer of facial expressions from a source video to a target video while preserving the target person's identity. Similarly, FaceSwap involves transferring the facial region from one video to another.DeepFakes

is a way of changing faces in videos using computers. It's like creating a mask for someone's face and putting it on another person's face in a video. NeuralTextures (NT) utilizes learned neural textures of the target individual to reenact facial expressions through deferred neural rendering.

Some Deepfake detection methods makes use of biological features in detection such as eye blinking, eyebrow, ear , eye movement , mouth movement and heartbeat detection[7]. This paper focuses on utilizing rPPG signals, which were previously utilized for heart rate estimation, for deepfake detection[4].

rPPG technology is based on the principle that under specific lighting conditions, changes in skin color within a fixed area can be captured via a camera. Analyzing these periodic pulsations allows for the extraction of physiological indicators such as heart rate, respiratory rate, and heart rate variability from the recorded skin tone information in the video [9]. Periodic heartbeats cause an increase in blood flow, resulting in some periodic changes in capillary volume, that is, changes in intravascular blood flow and haemoglobin content, which are usually described by the blood volume pulse (BVP). When light shines on the skin, it causes subtle changes that are hard to see without special equipment, but color sensors can detect these changes because they are very sensitive and can capture fine details. [9].

Chapter 2

Literature Review

2.1 Deepfake Creation

Deepfake techniques use advanced algorithms to create realistic fake content, including videos, images, text, and voices. Generative Adversarial Networks (GANs) are commonly used to generate forged images and videos. Deep autoencoders are a popular model in deep networks. They consist of two symmetrical Deep Belief Networks (DBNs), with several layers representing the encoding stage and the remaining layers representing the decoding stage [6]. The first success of deepfake technology was the development of FakeApp, which allowed one person's face to be swapped with another. The "FakeApp" software requires large amounts of data to produce better results. The data is passed to the system to train the model, which then inserts the face into the target video. Creating fake videos in FakeApp involves extracting all the images from the source video into a folder, cropping and aligning them correctly, and then processing them using a trained model. After merging the faces, the final video is ready [6].

2.2 Deepfake Detection

Early deepfake detection methods relied on hand-crafted features such as color histograms, which were manually designed to capture specific video content aspects. These features were then inputted into machine learning algorithms like SVMs or random forests for classification. However, with the emergence of deep learning, Convolutional Neural Networks (CNNs) became the preferred choice for deepfake detection. Initial CNN models like Mesonet and CapsuleNet were developed to detect crucial facial features in both original and deepfake videos.

The hypothesis that deepfake videos exhibit spatial and temporal consistencies drove the design of detection models. Various models were created to calculate local spatial inconsistencies, such as Multi-scale Patch Similarity, which measures pixel-wise differences between different areas. For temporal analysis, Recurrent Neural Networks (RNNs) were employed to explore temporal dependencies and patterns in video frames. Subsequently, the Attention mechanism was developed to enhance RNN models by assigning weights to each frame based on its relevance to deepfake detection .

Another approach used for deepfake detection is by analyzing biological features as discussed in [7]. Physiological measurements, including eyebrow recognition, eye blinking, eye movement, ear and mouth detection, and heart rate, can be used as biological signals to detect deepfakes.

Deepfakes lack natural eye blinking because they are generated by AI algorithms that often do not include realistic blinking patterns.The DeepVision [5] algorithm analyzes the human eye blinking pattern to detect deepfakes, focusing on detecting anomalies such as rapid and repeated eye blinking within a short period. They developed a method that combines Convolutional Neural

Networks (CNN) and Long-term Recurrent Convolutional Neural Networks (LRCN) to classify when eyes are open or closed, considering past patterns of eye behavior. By aligning faces and focusing on eye regions, the model effectively detects eye blinking in videos, showing promising results in identifying deepfake videos compared to standard datasets .

The creation of face-swap and lip-sync deepfakes often overlooks the human ear, which provides both static biometric signals and dynamic cues from jaw movement.. While face-swaps may accurately depict a co-opted identity, the ears typically belong to the original person, and in lip-sync deepfakes, ear dynamics are not synchronized with mouth and jaw movements. [1] describes a forensic technique leveraging these static and dynamic aural properties for deepfake detection.

Heartbeat signals, detected through photoplethysmography (rPPG) technology, have been utilized in biomedicine to monitor heart rate. Remote PPG (rPPG) technology enables the capture of subtle changes in skin color from recorded videos, which is disrupted by facial pixel modifications in deepfake videos. Previous studies have demonstrated that analyzing the rPPG signal can effectively detect deepfake videos, as deepfakes struggle to maintain consistent heartbeat signals, providing a reliable indicator for forgery detection.

[8] adopts a Multi-scale Spatial-Temporal PPG map for detection of heartbeat signals across multiple facial regions. It proposes a two-stage network comprising a Mask-Guided Local Attention module (MLA) to capture distinct local patterns of PPG maps and a Temporal Transformer to facilitate interactions between features of neighboring PPG maps over long distances.

[3] proposed an approach that not only distinguishes deepfakes from real videos but also identifies the specific generative model used to create the deepfake. Existing CNN-based methods learn the residuals of the generator

which contain valuable information that can be uncovered by disentangling them with biological signals. By observing that spatiotemporal patterns in biological signals can be seen as a projection of residuals, they extract PPG cells from real and fake videos and employ a state-of-the-art classification network to detect the generative model used in each video.

Chapter 3

Research Gaps and Problem Statement

3.1 Research Gap

In the realm of digital media forensics, the emergence and proliferation of deepfake technology have introduced a significant challenge to the authenticity and integrity of visual content. While substantial efforts have been directed towards detecting deepfake images, there exists a notable research gap in effectively utilizing Remote Photoplethysmography (rPPG) signals for detecting deepfake videos. Current methodologies primarily focus on analyzing uncompressed video data, disregarding the prevalent use of compressed videos on social media platforms. Compression algorithms employed by these platforms, such as spatial and temporal compression, alongside advanced video codecs like H.264 or H.265, significantly alter the visual characteristics of videos, complicating traditional detection techniques. Prior research in deepfake detection has often relied on intricate features such as edge detection, eye blinking patterns, and color imbalances, which pose challenges when videos undergo compression. The integration of audio signals alongside rPPG signals remains largely unexplored, despite the potential synergy between audio and

visual cues in enhancing the accuracy and robustness of deepfake detection systems. Therefore, there is a pressing need to develop tailored approaches that accommodate compressed video formats and exploit the interconnected nature of audio-visual cues for more effective deepfake detection.

3.2 Problem Statement

The proliferation of deepfake technology presents a pressing challenge to the authenticity and reliability of digital media content. While existing research has made significant strides in detecting deepfake images, the detection of deepfake videos remains a challenging task, particularly in the context of compressed video data prevalent on social media platforms. Current methodologies often fail to effectively utilize Remote Photoplethysmography (rPPG) signals for video-based deepfake detection, as they primarily focus on uncompressed video data. Compression algorithms employed by social media platforms alter the visual characteristics of videos, making traditional detection techniques inadequate. Furthermore, existing approaches heavily rely on intricate visual features that are difficult to interpret when videos undergo compression. Moreover, the potential synergy between audio and visual cues remains largely untapped in the context of deepfake detection. Therefore, the primary objective of this study is to address these challenges by proposing a novel framework that leverages rPPG signals extracted from both compressed video data and accompanying audio streams. By synthesizing information from multiple modalities, including visual, physiological, and auditory cues, this research aims to discern subtle discrepancies indicative of deepfake manipulation, thus advancing the state-of-the-art in deepfake detection and bolstering trust and reliability in digital media content.

Chapter 4

Proposed Methodology/ Solution

4.1 Data Preprocessing

The initial 300 frames of the video were extracted and processed. For each frame, facial regions were extracted, and 81 facial landmarks were computed using the DLIB library. These landmarks were used to identify key facial features, enabling the subsequent removal of eyes and background components based on the landmark information.

Subsequently, face alignment was performed for each detected face. This involved rotating the facial images to align them based on the angles of the eyes ensuring consistent orientation and positioning across all facial images.

A motion magnification algorithm was then applied to the processed facial images. This algorithm acts as a visual motion microscope, amplifying subtle motions within the video sequence enabling the visualization of minute deformations that would otherwise remain imperceptible.

Facial regions were divided into N rectangular blocks and average pooling was independently applied to each block across each color channel for every frame . This yields to the formation of rPPG (Remote photoplethysmography) map. Each row of map corresponds to temporal variation of one block across RBG channel .

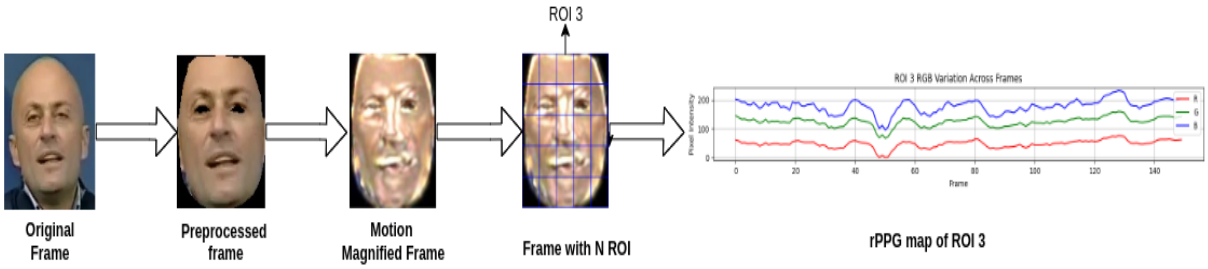


Figure 4.1: Stages of Preprocessing

4.2 Spatial and Temporal attention

The rPPG map can directly be used to classify the video as real and fake as it inherently contains the spatial and temporal information. However the contribution of different position in rPPG map are not always same due to effect of head movement, illumination variation , and sensor noises . Attention weights were therefore introduced to dynamically emphasize different regions of interest within the rPPG map .

Two types of attention weights were design , spatial (s) and temporal (t). The spatial attention higlights different block (ROI) to adapt the environment variation .For temporal attention two types of weights are calculated (t1 , t2) . For t1 we train an LSTM to which each row of rPPG map is sequentially filled to show temporal variation of face . For t2 pretrained model - MesoNet is used which was initially trained to calculate the fakeness of an image. Each frame is given input to MesoNet classifier and fakeness score was calculated. The frame having higher fakeness score contribute more to the final classification. Final temporal weight will be t = t1 + t2 .

4.3 Classification Model

For classification ResNet18 model is used.Cross-entropy loss function is used with Adam optimizer. The model has been compiled for 25 epochs and

32 batch size to master the training dataset . The training dataset contains videos from the FaceForensics++ dataset, comprising 300 original videos and 300 DeepFake videos.

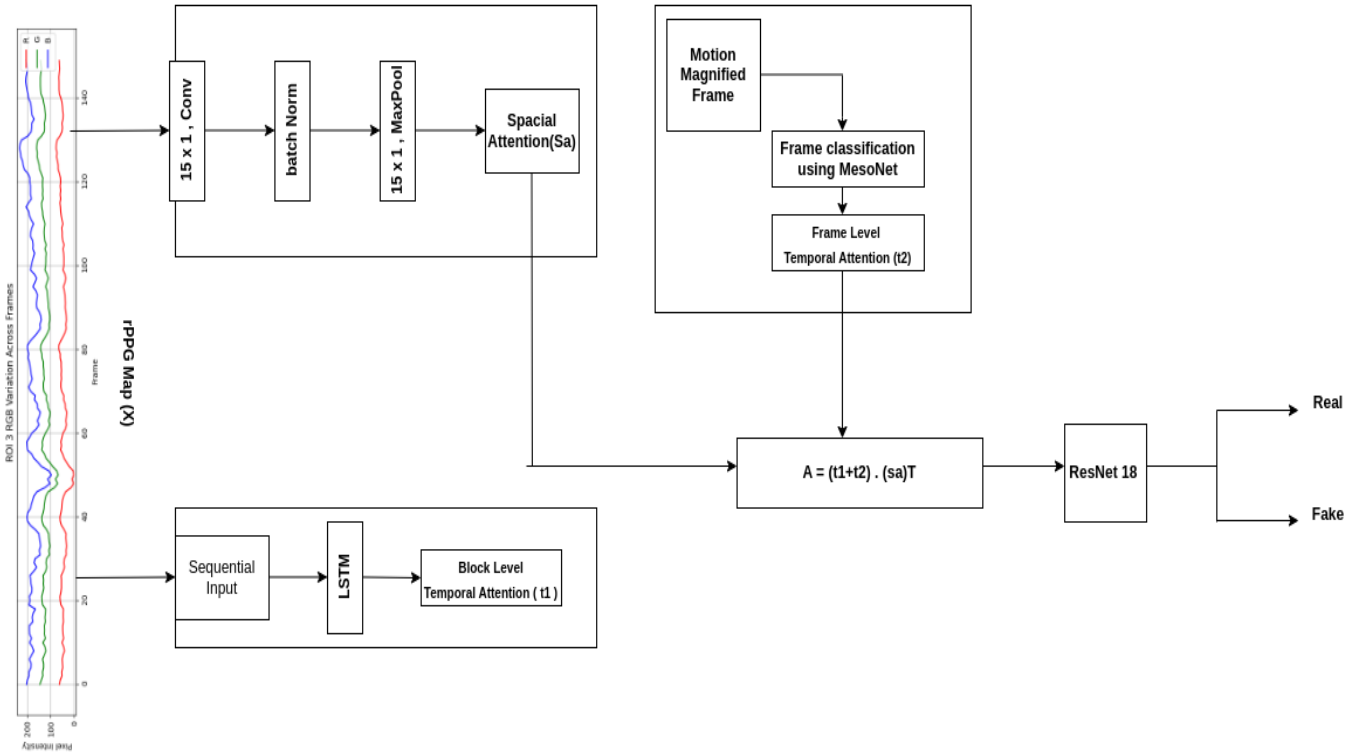


Figure 4.2: Model

Chapter 5

Results and Discussion

Confusion matrix and accuracy are the matrics used for the evaluation,

5.1 Confusion Matrix

A confusion matrix is a matrix that summarizes the performance of a machine learning model on a set of test data. It is composed of four main properties used to define the classifier's measuring metrics . The results of a classification algorithm are presented in a confusion matrix, depicted in Figure 6.1.

- True Positive (TP): The data is expected to be positive and is in fact positive.
- False Positive (FP): The data is expected to be positive but turned out to be negative.
- True Negative (TN): The data is expected to be negative and is in fact negative.
- False Negative (FN): The data is expected to be negative but turned out to be positive.

Serial No	Measures	Values
1	True Positive	13
2	False Positive	12
3	False Negative	6
4	True Negative	19

Table 5.1: Confusion Matrix

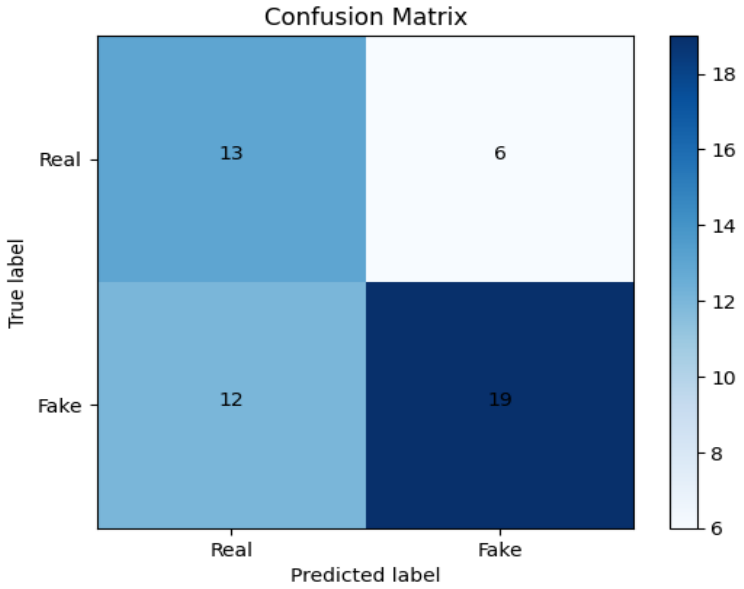


Figure 5.1: Confusion Matrix

5.2 Accuracy

Accuracy is a measure of the overall correctness of the classification system and is calculated as the ratio of correctly classified instances to the total instances. Accuracy = (TP+TN) / (TP+FN+FP+TN). The model was tested for 25 epochs with a batch size of 32 . It achieved a training accuracy of 73.21% , validation accuracy of 52.60% and a testing accuracy of 64%. The graphs 6.2 and 6.3 shown below illustrates these results. 50 videos were used for testing .

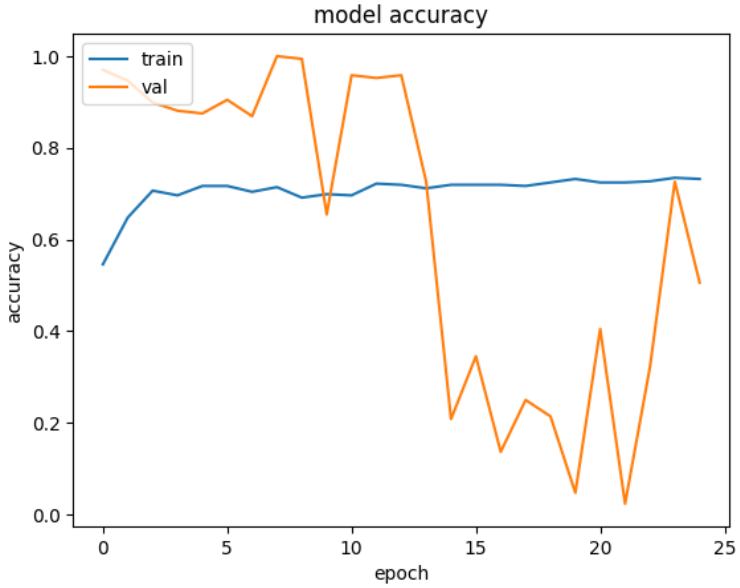


Figure 5.2: Accuracy Graph

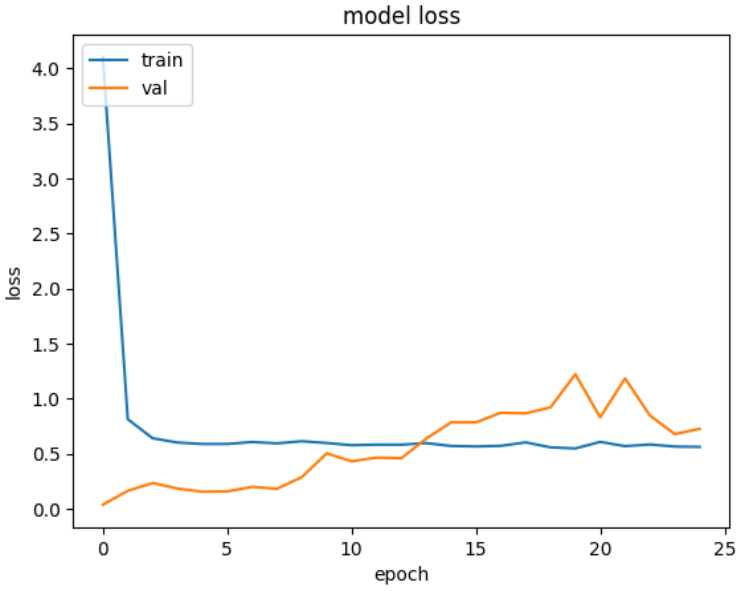


Figure 5.3: Loss graph

Chapter 6

Conclusion

In our project, we utilized the Remote photoplethysmography (rPPG) method along with spatial and temporal attention to classify videos as either fake or real. Our training dataset comprised 600 videos from the FaceForensics++ dataset, evenly split between 300 real and 300 fake videos of 'c23' compression quality, indicating high-quality compression using the H.264 codec. Despite achieving a training accuracy of 73.21%, we acknowledge that this result could be improved by expanding the dataset and fine tuning the hyperparameters like epochs and learning rate .

We intend to enhance our model's performance by training it on highly compressed videos (c40 quality) and datasets such as DFDC (Deepfake Detection Challenge). Furthermore, we aim to enhance the model's robustness, particularly in detecting deepfakes in highly compressed videos.

Bibliography

[1] Shruti Agarwal and Hany Farid. Detecting deep-fake videos from aural and oral dynamics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 981–989, 2021.

[2] Giuseppe Boccignone, Sathya Bursic, Vittorio Cuculo, Alessandro D'Amelio, Giuliano Grossi, Raffaella Lanzarotti, and Sabrina Patania. Deepfakes have no heart: A simple rppg-based method to reveal fake videos. In *International Conference on Image Analysis and Processing*, pages 186–195. Springer, 2022.

[3] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. How do the hearts of deep fakes beat? deep fake source detection via interpreting residuals with biological signals. In *2020 IEEE international joint conference on biometrics (IJCB)*, pages 1–10. IEEE, 2020.

[4] Javier Hernandez-Ortega, Ruben Tolosana, Julian Fierrez, and Aythami Morales. Deepfakeson-phys: Deepfakes detection based on heart rate estimation. *arXiv preprint arXiv:2010.00400*, 2020.

[5] Tackhyun Jung, Sangwon Kim, and Keecheon Kim. Deepvision: Deepfakes detection using human eye blinking pattern. *IEEE Access*, 8:83144–83154, 2020.

[6] Bahar Uddin Mahmud and Afsana Sharmin. Deep insights of deepfake technology: A review. *arXiv preprint arXiv:2105.00192*, 2021.

[7] Kundan Patil, Shrushti Kale, Jaivanti Dhokey, and Abhishek Gulhane. Deepfake detection using biological features: a survey. *arXiv preprint arXiv:2301.05819*, 2023.

[8] Jiahui Wu, Yu Zhu, Xiaoben Jiang, Yatong Liu, and Jiajun Lin. Local attention and long-distance interaction of rppg for deepfake detection. *The Visual Computer*, 40(2):1083–1094, 2024.

[9] Yuezheng Xu, Ru Zhang, Cheng Yang, Yana Zhang, Zhen Yang, and Jianyi Liu. New advances in remote heart rate estimation and its application to deepfake detection. In *2021 International Conference on Culture-oriented Science & Technology (ICCST)*, pages 387–392. IEEE, 2021.