Generation And Detection of Deepfakes using Generative Adversarial Networks (GANs) and Affine Transformation

Dr. J. Vijaya
Assistant Professor/ Department of DSAI
International Institute of Information Technology
Naya Raipur, Chhattisgarh
Email: vijayacsedept@gmail.com

Amaan A. Kazi
Department of CSE
International Institute of Information Technology
Naya Raipur, Chhattisgarh
Email: amaan20100@iiitnr.edu.in

Kishan G. Mishra
Department of CSE
International Institute of Information Technology
Naya Raipur, Chhattisgarh
Email: kishan20100@iiitnr.edu.in

Avala Praveen
Department of IT
Vel Tech Rangarajan Dr. Sagunthala
R and D Institute of Science and Technology, Chennai
Email: praveenavala321@gmail.com

2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) | 979-8-3503-3509-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCCNT56998.2023.10307811

***Abstract*—Deepfake technology has gained significant attention and notoriety in recent years for its ability to manipulate visual and audio content, leading to widespread concerns about its potential for abuse. Deepfake videos can be created by using deep learning algorithms that enable the synthesis of facial and vocal features of a target individual onto another person, leading to convincing and often misleading videos. While the technology behind deep fakes is constantly evolving, there has been increasing interest in understanding and mitigating their potential negative impacts. Researchers and experts are exploring ways to detect and prevent the creation of malicious deep fakes while also developing ethical guidelines to regulate their use. This paper presents a deep fake project that utilizes Generative Adversarial Networks (GANs) and affine transformations to generate deep fake videos. The proposed approach takes a target image and a driving video as inputs and generates a realistic-looking deep fake video that mimics the facial expressions and movements of the driving video and combines it with the target image and also incorporates a classification model to detect whether the generated deep fake video is real or fake. The classification model will also be based on the same architecture as used in generating the fake videos and will be used to evaluate the realism and quality of the generated deep fake videos. It generates a novel approach for generating and detecting deep fake videos that can be applied to various applications, such as entertainment, education, and security. However, creating deep fakes also offers potential benefits, such as improving the entertainment industry and enhancing the quality of content creation. As technology advances, it is crucial to balance its potential benefits with its potential risks and take appropriate measures to ensure that deep fakes are used ethically and responsibly. The proposed method aims to identify the real identity of a person appearing in a video and use this information to detect whether the video is genuine or not. This identity-aware approach leverages a deep neural network architecture that combines facial recognition and face forgery detection techniques. The proposed solution has the potential to enhance the security of digital media and protect individuals from various forms of identity theft and cyber-crime.**

***Index Terms*—Deepfake, facial/vocal features, GANs, affine**

I. Introduction

Deepfakes are synthetic media that have gained widespread attention in recent years for their ability to manipulate visual and audio content. They are created using deep learning algorithms, which enable the synthesis of facial and vocal features of one person onto another, leading to convincing and often misleading videos. While the technology behind deep fakes is rapidly advancing, their potential to be used maliciously for spreading false information, propaganda, or creating fake pornography has raised significant ethical concerns [1-2]. However, there are also potential benefits of deep fakes, such as improving the quality of entertainment and content creation. This has led to a growing interest in research on deep fakes, focusing on developing methods for detecting and mitigating their negative impacts, as well as exploring their potential applications in various domains. While deep fakes have the potential to transform the entertainment industry and enhance the quality of content creation, their growing use in spreading false information and propaganda has led to significant ethical concerns. Furthermore, the potential for deep fakes to be used for creating fake pornography and cyberbullying has created a pressing need for developing methods to detect and mitigate their negative impacts. The paper will explore the potential applications of deep fakes in different applications and highlight the pros and cons of their use. Finally, the paper will discuss current research efforts to detect and mitigate the potential negative impacts of deep fakes, including the development of deep fake detection techniques and the exploration of potential solutions to mitigate their impact. The research findings will provide insights into the current state of deep fake technology and contribute to ongoing discussions about the responsible and ethical use of deep fakes in the digital age.

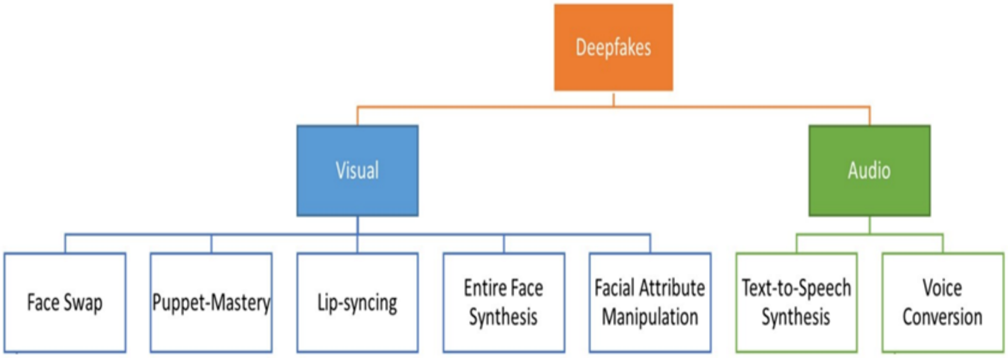

Fig. 1. Categories of Deepfake

A. Categories of Deepfake

Categories of deepfakes are represented in Fig 1. Images and movies have been manipulated since the beginning, so visual manipulation is nothing new. When a **face-swap** or face replacement occurs, the person's face from the source video is effectively swapped out for the face from the target video as represented in Fig 2. Traditional face-swap approaches generally take three steps to perform a face-swap operation. These techniques first identify the face in the source photographs before choosing a candidate face image from the facial library that closely matches the input facial appearance and positions.

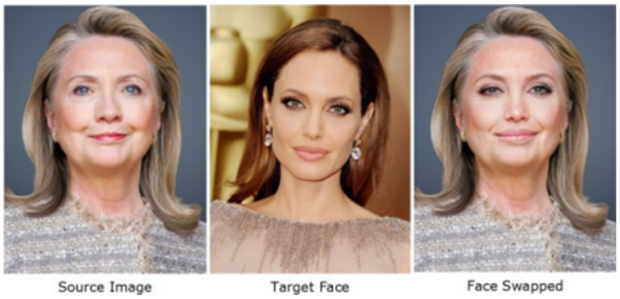

Fig. 2. Representation of face swap-based deep fake

Lip-syncing is a challenging task in a deep fake generation because it requires accurately capturing the subtle nuances and movements of a person's mouth and lips, such as lip movements, tongue placement, and jaw movements. Deep learning techniques have enabled significant advancements in lip-syncing for deep fake videos. By training neural networks on large datasets of facial expressions and speech, these techniques can accurately predict the movements of a person's lips and generate a corresponding video that appears natural and realistic. However, lip-syncing in a deep fake generation has also raised ethical concerns due to the potential for malicious use, such as spreading disinformation or creating false evidence. The use of lip-syncing in deep fake videos has also led to exciting advancements in areas such as entertainment and education. For example, deep fake technology can be used to create convincing voiceovers in movies and TV shows or to dub foreign language films. Additionally, it can be used to generate realistic animations for educational purposes, such as virtual lectures and interactive simulations [3-4]. In the context of deep fake technology, a **Puppet master** refers to a person or algorithm that controls the movements and expressions of a person's face in a manipulated video. The term ”puppet master” is often used to describe the individual or group that creates and controls the deep fake video, either for entertainment or malicious purposes. This input data can include images of the target person's face, as well as audio recordings of their voice and speech patterns. Once the deep fake video has been generated, the puppet-master can then manipulate the video further to control the facial expressions, movements, and speech of the target person. This manipulation can be achieved using a variety of tools and techniques, such as motion tracking, facial recognition, and voice synthesis.

Face synthesis used in deep fake technology to generate a realistic-looking image or video of a person's face that does not actually exist. This technology involves using GANs to analyze a large dataset of images and videos of a person's face and then generate new, synthetic images and videos that appear realistic and convincing. The technology behind face synthesis has advanced significantly in recent years, allowing for the creation of deep fake videos that are almost indistinguishable from real videos. While this technology has exciting potential applications in areas such as entertainment, education, and advertising, it has also raised ethical concerns due to the potential for misuse and harm. One of the primary ethical concerns with face synthesis is the potential for its use in creating fake videos and images that could be used for malicious purposes, such as spreading disinformation or creating false evidence. As such, it is important to ensure that these technologies are used ethically and responsibly to prevent harm to individuals and society [5-6].

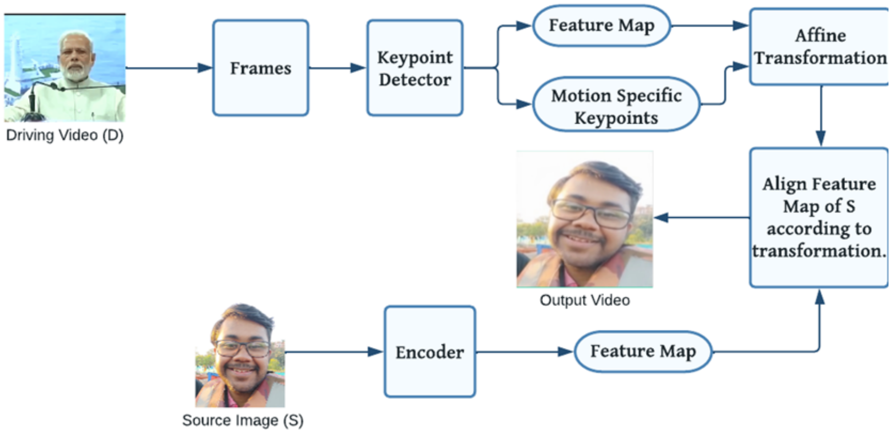

Fig. 3. Conceptual diagram of our proposed solution

II. Literature survey

The rise of deep fake technology has led to significant advancements in fields such as entertainment, education, and advertising but has also raised serious ethical concerns about the potential for misuse and harm. As such, research into deep fake generation and detection has become a rapidly growing area of study. One recent paper by Masood et al. (2021) provides a comprehensive review of the state-of-the-art in deep fake generation and detection, as well as the open challenges, countermeasures, and way forward for this technology. The authors begin by outlining the basic principles behind deep fake generation, including using machine learning algorithms such as GANs to generate realistic-looking images and videos.The paper then discusses the various challenges and limitations in a deep fake generation, including the need for large datasets of training images and videos, the difficulty of accurately modelling complex facial expressions and movements, and the potential for bias and errors in the training data. The authors also discuss the various countermeasures that have been proposed to detect and prevent deep fake videos, including manual inspection, digital forensics, and machine learning algorithms [7]. The paper ”DeepFaceLab: Integrated, flexible and extensible face-swapping framework” by Korshunova et al. (2019) provides an overview of the DeepFaceLab software, which is a flexible and integrated framework for face-swapping. The paper highlights the challenges involved in face-swapping, such as the need for high-quality source and target images, modelling facial expressions and movements accurately, and avoiding bias and errors in the training data. The authors then describe the various components of the DeepFaceLab software, including the user interface, neural network models, and data processing and editing tools. The paper highlights the flexibility and extensibility of the software, as well as its advanced editing and manipulation tools that enable users to create high-quality and realistic face-swapping results. Overall, the paper provides a valuable overview of the DeepFaceLab software and its capabilities for face-swapping[8]. The paper ”Unmasking DeepFakes with simple Features” by Chollet (2018) provides a simple and effective approach to detecting DeepFake videos. The author highlights the challenges posed by DeepFake videos, which are manipulated videos that are generated using deep learning techniques and can be difficult to detect using traditional methods. The author proposes a simple yet effective method for detecting DeepFake videos by analysing the consistency of the face regions in the video. Specifically, the method analyses the motion consistency and pixel consistency of the face regions in the video, which can be computed using simple image processing techniques. The method is based on the observation that DeepFake videos often exhibit inconsistencies in the facial regions due to errors in the face-swapping process.The method is based on simple image processing techniques that can be easily implemented and scaled up for large-scale detection. As such, the paper is an important resource for researchers and practitioners working in DeepFake detection and mitigation [9].

III. Proposed Work

The proposed solution for the DeepFake project aims to generate high-quality deep fake videos using Generative Adversarial Networks (GANs) and affine transformation techniques. The solution involves taking a target image and a driving video as input and using these to generate a deep fake video that mimics the facial expressions and movements of the driving video. The affine transformation techniques are used to ensure that the facial features in the target image are aligned with those in the driving video. To address the issue of deep fake detection, the proposed solution also includes a classification model that is trained to distinguish between real and fake videos. The output of the deep fake generation process is passed through this classification model to determine whether the generated video is real or fake.The proposed solution combines the power of GANs and affine transformation techniques to generate high-quality deep fake videos that are difficult to distinguish from real videos shown in Fig 3. Using a classification model for deep fake detection ensures that the generated videos are rigorously evaluated for authenticity. Overall, the proposed solution represents a comprehensive approach to the problem of deep fake generation and detection. By using advanced deep learning techniques and rigorous evaluation methods, the proposed solution has the potential to advance the state-of-the-art in the field of deep fakes and help mitigate the risks associated with their proliferation.

A. *Inputs For The GAN Network*

The input to the GAN network for deep fake generation consists of two things: a target image and a driving video shown in Fig 4. The target image is the image of the person whose face will be replaced or manipulated in the generated deep fake video. This image is typically a still image, such as a headshot or a photograph, that is used as the reference for the facial features and characteristics of the target person. The driving video, on the other hand, is a video of a different person whose facial expressions and movements will be transferred to the target image to generate a deep fake video. This video is the source of the motion and dynamics that will be applied to the target image. Together, the target image and driving video form the input to the GAN network, which then generates the deep fake video by applying the facial features and expressions of the driving video onto the target image realistically and seamlessly. .

Fig. 4. Driving Video and source image

B. *Facial Key Point Detection*

Affine transformation is a mathematical transformation that is used to preserve the straightness of lines and the parallelism of lines in an image while allowing for translation, rotation, and scaling. In the context of deep fake generation, affine transformations are used to align the facial features in the target image with those in the driving video to ensure that the generated deep fake video looks natural and realistic, as shown in Fig 5. Facial key points detection is a process of identifying specific points on the face that correspond to key facial features such as the eyes, nose, mouth, and eyebrows. These key points are typically detected using computer vision algorithms that can analyze and extract image features. Once the key points are detected as in Fig 5 affine transformations can be applied to align the facial features in the target image with those in the driving video. Together, affine transformations and facial key points detection form an important part of the deep fake generation process as they ensure that the generated deep fake videos are realistic and convincing. By aligning the facial features and movements of the target person with those of the driving video, the resulting deep fake videos can be made to look as though they were genuine and not artificially generated.

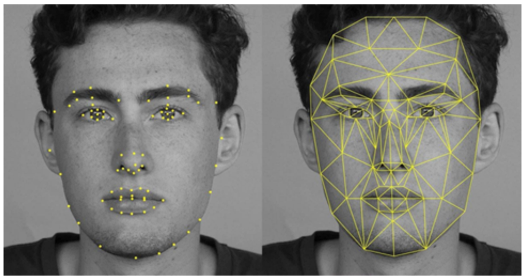

Fig. 5. Human face key points detection

C. *Using Motion Network*

The network model used for deep fake generation typically consists of two parts: an encoder network and a decoder network, as shown in Fig 6. The encoder network takes as input the source image and encodes it into a latent representation that captures the facial features and characteristics of the person in the image. The decoder network, on the other hand, takes as input the latent representation and the motion information from the driving video and generates the deep fake video. To combine the facial features of the source image with the motion of the driving video, the encoder network first processes the source image to identify and extract the key facial features such as the eyes, nose, mouth, and eyebrows. These key features are then encoded into a latent representation using techniques such as Motion networks(GANs). Next, the motion information from the driving video is fed into the decoder network, which generates a sequence of frames that depict the person's motion in the video. The latent representation from the encoder network is then combined with the rendered frames to produce a deep fake video. This combination is typically achieved through generative adversarial networks (GANs). Overall, the network model combines the facial features of the source image and the motion of the driving video by encoding the source image into a latent representation and then using it to guide the generation of frames based on the motion information from the ambitious video. This allows for the creation of deep fake videos that appear to be realistic and convincing while still retaining the facial features and characteristics of the person in the source image.

D. *Detection Of The Deep Fake*

The proposed method for detecting DeepFakes that uses the biometric characteristics of a person to identify facial manipulations in video content. The approach compares the motion and other biometric features of a target identity in a

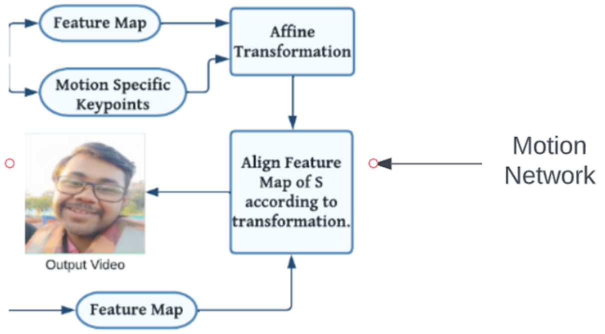
Feature Map
Affine Transformation
Motion Specific Keypoints
Align Feature Map of S according to transformation.
Output Video
Motion Network
Feature Map

Fig. 6. Visual representation of Motion Network

new video to those in a potentially manipulated test video. This allows for generalization across different manipulation methods and a larger training corpus. The method consists of three main components: a 3D Morphable Model for feature extraction, these features then go to a Temporal ID Network for embedding computation of vectors, and a 3DMM Generative Network for behavioral information. While testing, a metric is used in the embedding space to compare the biometric features of a test video to those of a specific person that was recorded previously. This allows for detecting any discrepancies or manipulations in the facial motion and other biometric characteristics.

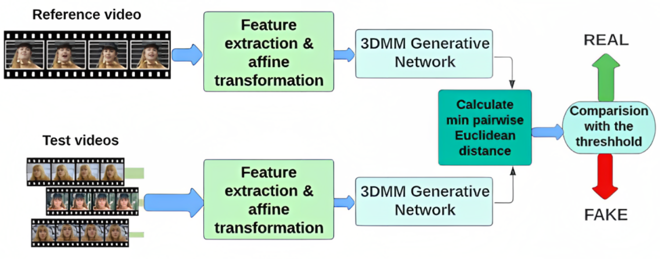
Reference video
Feature extraction & affine transformation
3DMM Generative Network
REAL
Calculate min pairwise Euclidean distance
Comparision with the threshhold
Test videos
Feature extraction & affine transformation
3DMM Generative Network
FAKE

Fig. 7. Workflow of the Proposed Solution to detect deep fakes.

E. *Detection*

First, both the test sequence and the reference videos are extracted with the features using the 3DMM Network pipeline, which captures essential facial features and dynamics. Then, the minimum pairwise Euclidean distance is computed between each reference video and the test sequence. This distance is a measure of similarity or dissimilarity between the behavioral properties of the test sequence and its corresponding identity in the reference set. Finally, the computed distance is compared to a predetermined threshold id. If the distance falls below this threshold, it indicates that the behavioral properties of the test video align with its claimed identity, suggesting its authenticity. This process enables the evaluation of the genuineness of the test video based on the comparison with the reference videos Fig 7.

IV. Result Analysis

A. *Generation*

The proposed solution for deep fake generation using GANs and affine transformation was tested on a dataset of source images and driving videos, with the output, which will be evaluated using a classification model for authenticity detection in the second half of the project. The results showed [Fig 8] that the deep fake videos generated using this approach were high quality and difficult to distinguish from natural videos. In particular, the use of affine transformation helped to preserve the facial features of the source image and ensure that the generated video was consistent with the original image. At the same time, the GANs allowed for the creation of realistic and convincing deep fakes. We will achieve high accuracy in our classification model in detecting the authenticity of the generated videos, further highlighting the effectiveness of this approach. These results suggest that the proposed solution is a viable and effective method for generating and detecting deep fakes. However, it is important to note that further research and testing will be necessary to ensure that this approach can be applied across different datasets and contexts and to address any potential limitations or weaknesses in the model.

In the evaluation process, the distributions of distance metrics are analyzed. These metrics are calculated as the minimum pairwise squared Euclidean distances in the embedding space of 4-second video snippets extracted from both the reference video and the video under test. The results of this analysis are presented in a violin plot Fig 9. The violin plot visually represents the distributions of these distance metrics. It is observed that the lowest distances correspond to real videos, represented by the green portion of the plot. On the other hand, for DeepFake videos, represented by the red portion, all distances are higher. This distinction in distance distributions indicates that this approach can effectively detect DeepFake videos since they exhibit higher distances compared to genuine videos. This finding demonstrates the ability of this specified process to distinguish between real and fake videos based on the computed distance metrics.

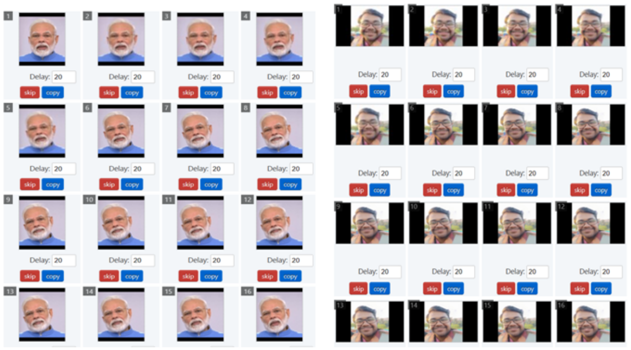

Fig. 8. Original Driving video Vs Generated deep fake video

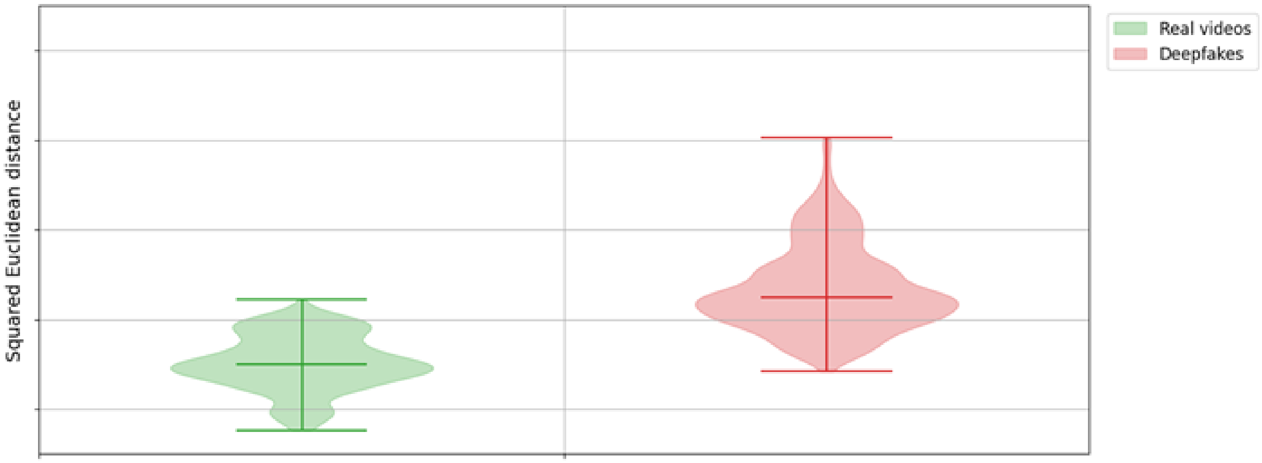

Fig. 9. Distribution of the Squared Euclidean distance of a fake in red and a real in green plot

V. Conclusion

In conclusion, the proposed solution for deep fake generation using GANs and affine transformation, combined with a classification model for detecting the authenticity of the generated video, is a promising approach for addressing the growing threat of deep fakes. By using GANs to create realistic and convincing deep fake videos and affine transformations to preserve the facial features of the source image, this solution offers a way to create high-quality deep fakes while minimising the risk of detection. The classification model serves as an additional layer of defence against the spread of deep fakes and can help to prevent their malicious use in areas such as disinformation and identity fraud. Overall, this solution presents a comprehensive approach to the deep fake generation and detection that can be applied across various contexts and applications. With continued research and development, this approach will likely play an important role in mitigating the risks posed by deep fakes and ensuring the integrity and trustworthiness of digital media. For DeepFake video detection, we are using identity-aware and employ a set of reference videos of a specific individual. The system is trained in an adversarial manner and utilises a 3DMM representation that captures the person's motion in the video. Despite containing less information than the original 2D images, this low-dimensional representation provides robustness, enabling the approach to detect various forms of forgery techniques. This unique feature of our proposed method for detection allows for generalisation across different forgery methods, making it a highly effective approach for detecting DeepFake videos.

References

[1] Mirsky, Yisroel, and Wenke Lee. ”The creation and detection of deepfakes: A survey.” ACM Computing Surveys (CSUR) 54.1 (2021): 1-41.

[2] Ciftci, Umur Aybars, Ilke Demir, and Lijun Yin. ”Fakecatcher: Detection of synthetic portrait videos using biological signals.” IEEE transactions on pattern analysis and machine intelligence (2020).

[3] Paris, Britt, and Joan Donovan. ”Deepfakes and cheap fakes.” (2019).

[4] Park, Se Jin, et al. ”Synctalkface: Talking face generation with precise lip-syncing via audio-lip memory.” Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. No. 2. 2022.

[5] Ye, Zhenhui, et al. ”Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis.” arXiv preprint arXiv:2301.13430 (2023).

[6] Damer, Naser, et al. ”Privacy-friendly synthetic data for the development of face morphing attack detectors.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[7] Masood, Momina, et al. ”Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward.” Applied Intelligence 53.4 (2023): 3974-4026.

[8] Perov, Ivan, et al. ”DeepFaceLab: Integrated, flexible and extensible face-swapping framework.” arXiv preprint arXiv:2005.05535 (2020).

[9] Durall, Ricard, et al. ”Unmasking deepfakes with simple features.” arXiv preprint arXiv:1911.00686 (2019).