DEEPFAKE DETECTION: What, How and Improvements

Indian Institute of Information Technology, Guwahati

Under the guidance of: Dr. Ferdous Ahmed Barbhuiya

Abhay Chaudhary

Roll no.: 2101005

`

IIIT GUWAHATI April 2024

Table of Contents

Declaration

I hereby declare that, except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted, in whole or in part, for consideration for any other degree or qualification in this or any other university.

Abhay Chaudhary

April 2024

Acknowledgement

I extend my sincere appreciation to the Department of Computer Science and Engineering for providing me with the opportunity to undertake a research project that resonates closely with my academic pursuits. Furthermore, I wish to convey my profound gratitude to my mentor, Dr. Ferdous Ahmed Barbhuiya, for his steadfast support, invaluable guidance, and unwavering motivation throughout the entirety of the project. Dr. Barbhuiya's mentorship proved indispensable, particularly during challenging phases, and I am profoundly grateful for his insightful direction.

Abstract

This report comprehensively investigates the domain of deepfakes, sophisticatedly manipulated media content generated by deep learning algorithms. Deepfakes pose significant challenges due to their potential misuse, including misinformation dissemination and privacy breaches. The study explores current techniques for detecting deepfakes, encompassing both traditional methods and advanced machine learning approaches. Additionally, it examines potential improvements to existing detection methods, such as leveraging adversarial training and multimodal analysis. Furthermore, the report addresses the limitations of current deepfake detection techniques, highlighting areas for future research and development to enhance detection accuracy and robustness in combating the proliferation of deceptive media content.

1 - Introduction to Deepfakes

Deepfakes, a portmanteau of "deep learning" and "fake," represent a paradigm shift in the manipulation of multimedia content. Leveraging advancements in deep learning, particularly generative models like generative adversarial networks (GANs) and autoencoders, deepfake technology enables the synthesis of highly realistic yet entirely artificial images, videos, and audio recordings.

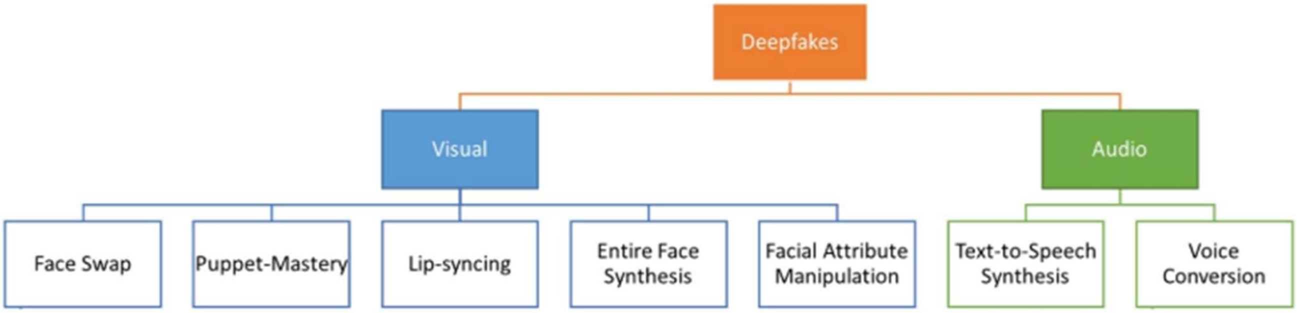
Initially gaining prominence through the creation of forged celebrity videos, deepfakes have evolved into a multifaceted challenge with profound societal implications. Their potential for malicious exploitation, including misinformation dissemination, identity fraud, and privacy violations, underscores the urgency of understanding and addressing this emerging phenomenon.

The proliferation of deepfake technology poses significant challenges across various domains, from journalism and entertainment to national security and personal privacy. The ability to manipulate audiovisual content with unprecedented fidelity blurs the line between truth and fiction, eroding trust in media and exacerbating existing concerns regarding the authenticity of digital content.

In response to the growing threat posed by deepfakes, researchers and practitioners have dedicated efforts to developing detection and mitigation techniques. However, the game between deepfake creators and detection algorithms continues, highlighting the need for ongoing research and innovation in this field.

2 - Components of a “Convincing” Deepfake

A convincing deepfake typically involves both its audio and video components for discrepancies or inconsistencies.



Audio discrepancies consist of:

- TEXT-TO-SPEECH-SYNTHESIS - Text-to-speech synthesis involves the conversion of written text into spoken words through computational algorithms, mimicking the cadence, intonation, and timbre of human speech. This process utilizes advanced neural networks to generate audio that closely resembles natural human speech patterns.
- VOICE CONVERSION- Voice conversion is a sophisticated technique that transforms the characteristics of one speaker's voice into those of another, allowing for the alteration of vocal attributes such as pitch, tone, and accent while preserving linguistic content.

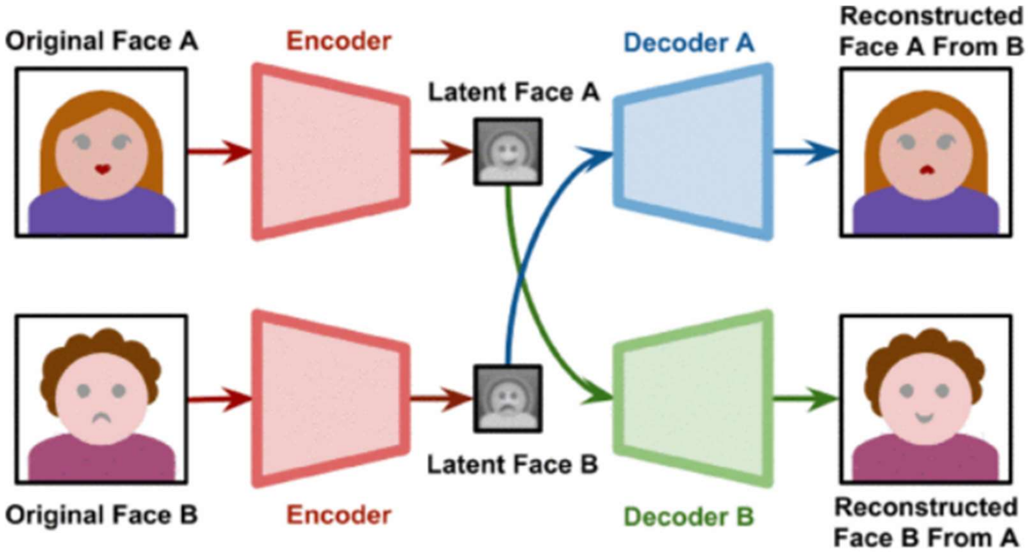
Video discrepancies consist of:

- FACE SWAP - Also known as facial reenactment or face replacement, involves the digital manipulation of facial features within an image or video, enabling the seamless substitution of one person's face with another. This process employs deep learning algorithms to analyze and manipulate facial landmarks, allowing for precise alignment and blending of facial expressions, skin textures, and lighting conditions.

- LIP SYNCING - Synchronizing the movements of a manipulated face with pre-existing audio, ensuring that the mouth movements match the spoken words convincingly.
- FACIAL ATTRIBUTE MANIPULATION - Facial attribute manipulation encompasses altering specific characteristics of a person's face in digital images or videos. This process involves using advanced algorithms to modify attributes such as age, gender, expression, or ethnicity while preserving the overall appearance and realism of the face.
- ENTIRE FACE SYNTHESIS - Entire face synthesis involves the creation of a completely new facial image that does not correspond to any existing individual.
- PUPPET MASTERY - This process allows creators to choreograph intricate movements, expressions, and interactions of virtual entities, imbuing them with lifelike behaviors and personalities.

3 – Deepfake Generation

Generative networks and encoder-decoder networks, particularly autoencoder-decoder models, are commonly used methodologies for Deepfake creation. In the encoder-decoder approach, original and target faces are encoded to create latent representations, allowing for the interchange of facial features between the two faces. The error function associated with the network determines the quality of the generated output, guiding adjustments to network weights for improved results. Generative Adversarial Networks (GANs) are another popular technique for Deepfake implementation, introduced by Ian Goodfellow and other researchers in 2014.



4- Open issues and future direction

Although great efforts have been made in devising deepfake generation and detection, there are several issues yet to be addressed successfully. In the following, some of them are discussed.

4.1. Generalization Capability

It is easy to notice in the literature that most of the existing deepfake detection frameworks' performances decrease remarkably when tested under deepfakes, manipulations, or databases that were not used for the training. Thus, detecting unknown novel deepfakes or deepfake generation tools is yet a big challenge. The generalization capability of deepfake detectors is vital for dependable precision and public trust in the information being shared online. Some preliminary generalization solutions have been proposed, but their ability to tackle novel emerging deepfakes is still an open issue.

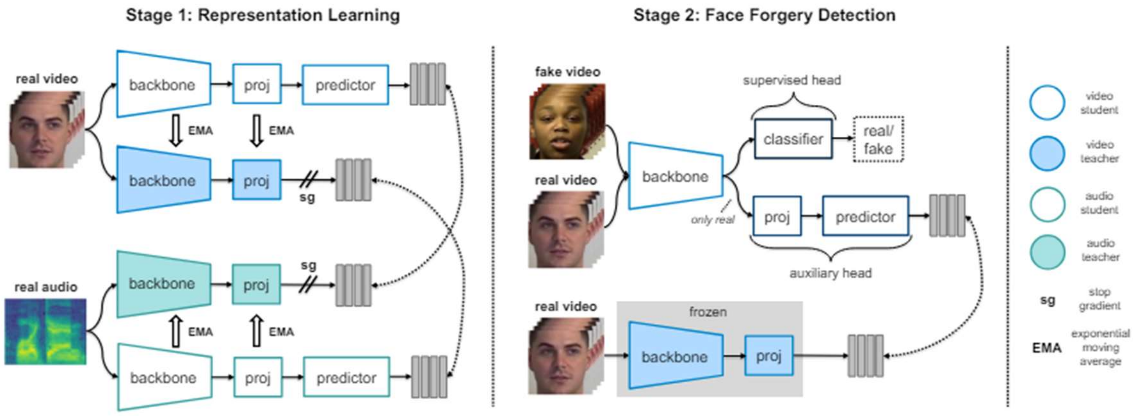
4.2. Explainability of Deepfake Detectors

There is a lack of work on the deepfake detection framework's interpretability and dependability. Most deep-learning-based deepfake or face manipulation detection methods in the literature usually do not explain the reason behind the final detection outcome. It is mainly due to deep learning techniques being the black box in nature. Current deepfake or face manipulation detectors only give a label, confidence percentage, or fakeness probability score but not the insight description of results.

4.3. Next-Generation Deepfake and Face Manipulation Generators

Improved deepfake and face manipulation generation techniques will help develop more advanced and generalized deepfake detection methods. Some of the shortcomings of current datasets and generation methods are the lack of ultra-high-resolution samples (e.g., existing methods are usually generating 1014 × 1024 resolution samples, which is not sufficient for the next generation of deepfakes), limited face attribution manipulations (i.e., face attribute manipulation types are dependent on the training set, thereby manipulation characteristics and attributes are limited, and novel attributes cannot be generated), video continuity problem (i.e., the deepfake/face manipulation, especially identity swap, techniques neglects the continuation of video frames as well as physiological signals), and no obvious deepfake/face manipulations (i.e., present databases are not composed of obvious fake samples such as a human face with three eyes).

5 –Deepfake Detection



The detection of deepfakes involves a multi-stage process, with each stage focusing on specific aspects of the manipulated content.

In the first stage, referred to as representation learning, temporally dense video representations are learned using cross-modal self-supervision from a large dataset of natural talking faces. This involves encoding real videos and their corresponding audio into embeddings that capture facial appearance and behavior information. A student-teacher framework is utilized, where teacher networks generate targets that student networks from the opposite modality must predict. Random masking techniques are applied to the inputs of the student networks to encourage contextual inference and prevent over-reliance on specific features.

In the second stage, termed multi-task forgery detection, the learned representations are leveraged for face forgery classification. This involves using a shared backbone network with supervised and auxiliary heads for forgery classification and target prediction, respectively. The video teacher from the first stage produces targets for the network to predict, while simultaneously performing forgery detection. The auxiliary loss encourages the network to focus on high-level spatiotemporal characteristics of facial appearance and behavior.

The overall objective is to minimize a combination of supervised and auxiliary losses, with the network being optimized via gradient descent. Throughout both stages, implementation details such as network architectures, optimization algorithms, and preprocessing techniques are carefully considered to ensure effective training and detection performance.

6- Contributions-

Contribution A- Implementing Explainable AI (XAI) techniques in a deepfake detection model, I collected results and emphasized discrepancies.

Model used for implementation:

XceptionNet

Dataset used:

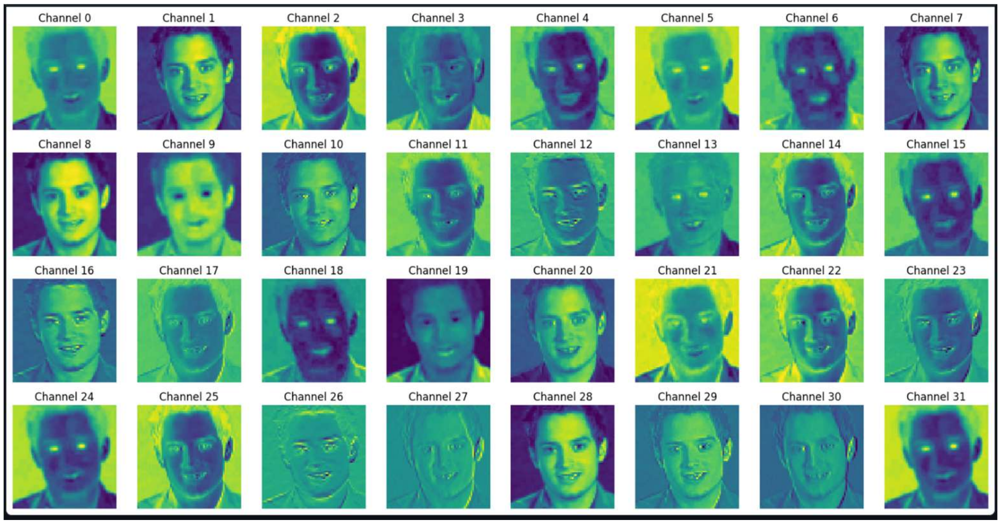
a. FaceForensic++
b. Celeb-DF

` **Technologies implemented-**

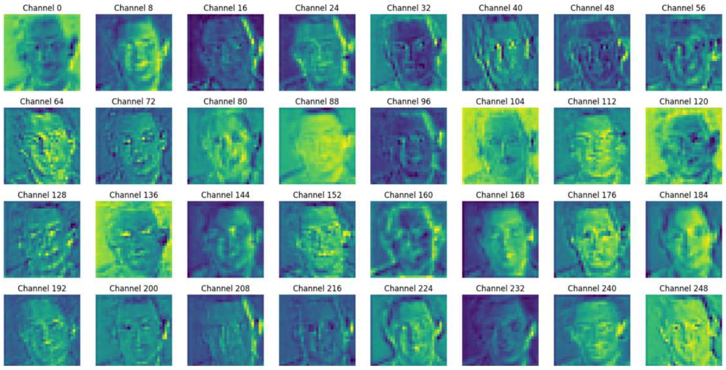
a. **GRAD-CAM**: Grad-CAM highlights the important regions of an image or feature map by visualizing the gradient of the target class with respect to the final convolutional layer.

OUTPUTS:

I. After one convolution layer

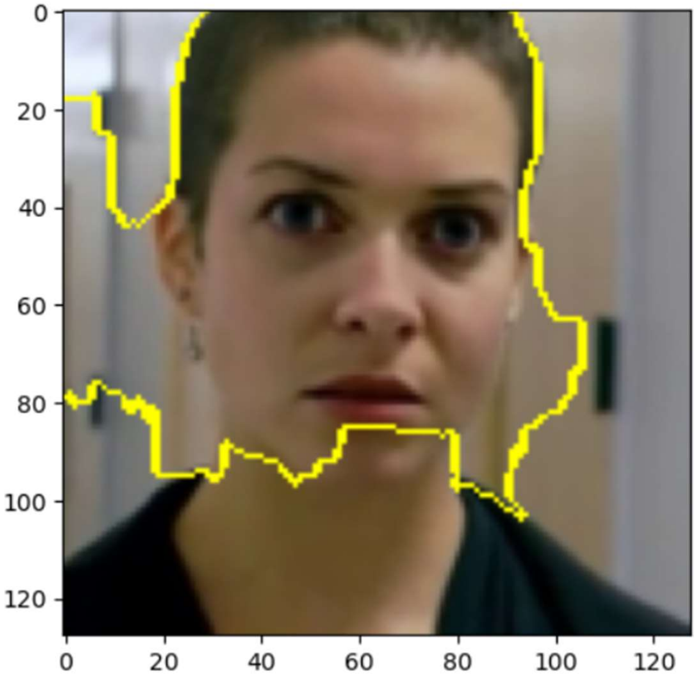


II. Separable Convolutional layer 2

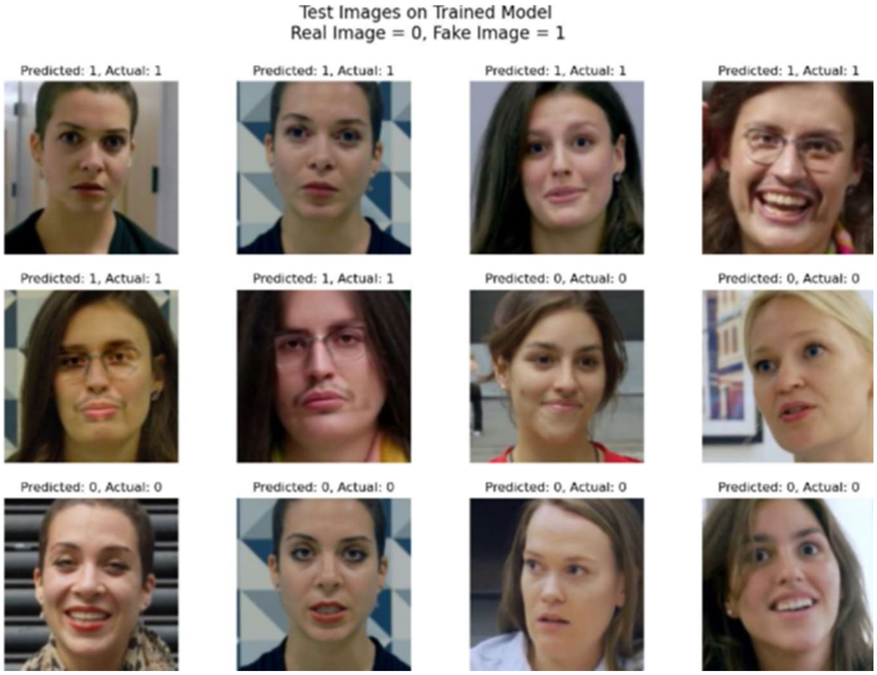


b. **LIME**: LIME (Local Interpretable Model-agnostic Explanations) explains the predictions of machine learning models by approximating their behavior locally, providing insights into how individual features contribute to specific predictions.

OUTPUT



c. **UPON PREDICTION:**



Contribution B-Drawing from my study of Mixed Generative Adversarial Networks (MGANs), I've devised a model primed for future implementation that's in place to counter 2 major issues.

Issues being-

1. The mode collapsing problem – Here, the generator fails to capture the diversity of the underlying data distribution, resulting in the production of limited or repetitive samples.

2. Enhancing the detection of distorted or blurry fake images through an enhancer layer of MGANs.

The concept centers on integrating a layer of Mixed Generative Adversarial Networks (MGAN) to enhance the clarity of blurry images, complemented by the detection segment from the REALFORENSICS model, offering a comprehensive solution for image refinement and fake image detection.

THE ARCHITECTURE FOR THE MODEL

1 - Mixed Generative Adversarial Networks (MGAN) Layer:

Consists of a generator network and a discriminator network.

The generator aims to enhance the clarity and quality of input images, particularly those that are blurry or distorted.

The discriminator evaluates the realism of the generated images to provide feedback to the generator.

2 - Feature Extraction Layer:

Extracts relevant features from the enhanced images generated by the MGAN layer.

These features are used as inputs for the detection segment.

3 - Detection Segment (from RealForensics Model):

Incorporates the detection mechanism from the REALFORENSICS (Fake Image Detection using Convolutional Neural Networks) model.

Utilizes convolutional neural networks (CNNs) or similar architectures to classify images as real or fake based on extracted features.

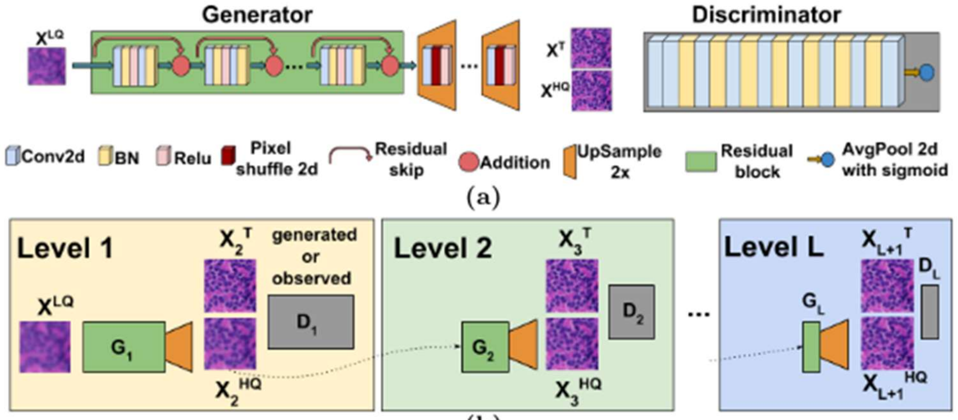
The detection segment provides feedback to the MGAN layer, helping to refine the image enhancement process based on the likelihood of the image being fake.

4 - Output Layer:

Produces the final output, which includes enhanced images with improved clarity and authenticity labels indicating whether the image is real or fake.

VISUAL REPRESENTATION

1-Mixed Generative Adversarial Networks (MGAN) Layer:



(a)

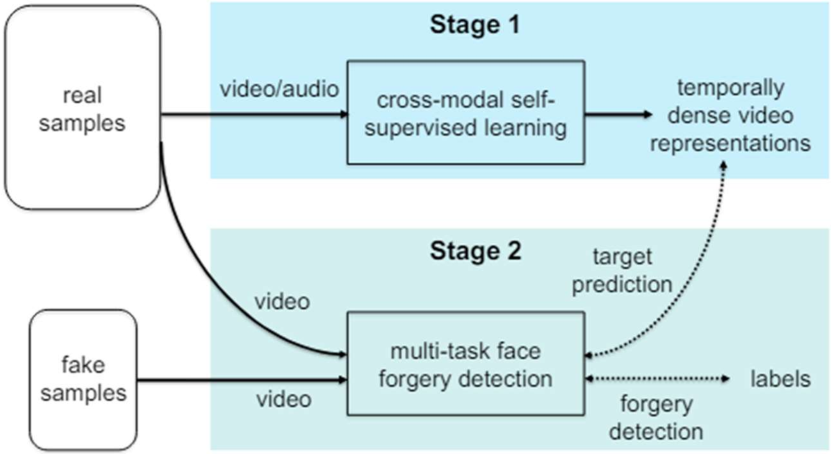
(b)

::::::::::::::::::::

2-Convolutional Layer responsible for feature extraction

::::::::::::::::::::

3- Detection Segment (from REALFORENSICS Model):


real samples
Stage 1
video/audio
cross-modal self-supervised learning
temporally dense video representations
Stage 2
target prediction
video
multi-task face forgery detection
fake samples
video
labels
forgery detection

::::::::::::::::::::

4-Output Layer: Feature Vectors --> Fake Image Detection Model --> Output Classification (Real/Fake)

7-CONCLUSION:

In conclusion, this research has not only advanced the field of image synthesis and fake image detection but has also integrated elements of Explainable Artificial Intelligence (XAI). By leveraging XAI techniques, such as feature visualization and saliency mapping, we have gained insights into the inner workings of our model, shedding light on how it makes decisions and providing interpretability to its outputs. This transparent approach not only enhances trust in the model's results but also opens avenues for further refinement and optimization. As we continue to explore the intersection of MGANs, detection mechanisms, and XAI, we move closer to creating robust and accountable AI systems that address real-world challenges with clarity and reliability.

REFERENCES:

1. https://arxiv.org/pdf/2311.01458v1
2. https://arxiv.org/pdf/2106.15575
3. https://arxiv.org/pdf/2201.07131
4. https://medium.com/arocketman/extracting-features-from-convolutional-neural-networks-for-image-retrieval-87b5127f8a92
5. https://openreview.net/pdf?id=rkmu5b0a-
6. https://ieeexplore.ieee.org/document/9544522
7. https://paperswithcode.com/paper/leveraging-real-talking-faces-via-self/review/?hl=111302
8. https://paperswithcode.com/sota/deepfake-detection-on-fakeavceleb-1
9. https://datacamp.com/tutorial/explainable-ai-understanding-and-trusting-machine-learning-mode
10. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9092227
11. https://arxiv.org/pdf/2105.05902.pdf