Human-Intelligent Systems Integration
https://doi.org/10.1007/s42454-024-00054-8

REVIEW ARTICLE

Exploring deepfake technology: creation, consequences and countermeasures

Sami Alanazi[1] · **Seemal Asif**[1]

Received: 13 February 2024 / Accepted: 17 August 2024

Abstract
This paper presents a comprehensive examination of deepfakes, exploring their creation, production and identification. Deepfakes are videos, images or audio that are remarkably realistic and generated using artificial intelligence algorithms. While they were initially intended for entertainment and commercial use, their harmful social consequences have become more evident over time. These technologies are now being misapplied for the creation of explicit content, coercing individuals and disseminating false information, resulting in an erosion of and potentially negative societal consequences. The paper also highlights the significance of legal regulations in controlling the utilization of deepfakes and investigates methods for their identification through machine learning. In the modern digital world, comprehending the ethical and legal implications of deepfakes necessitates a thorough understanding of the phenomenon.

Keywords Deepfake · GAN · False information · Detection methods · Image alteration · Counterfeit content · Realistic videos

1 Introduction

Deepfake images and videos are content that appears authentic, yet they are, in fact, created using artificial intelligence algorithms. Detecting such content can be challenging for the human eye as it is technically manipulated. Deepfakes are a blend of "deep learning" and "fake" videos which involve digitally altering videos to create hyper-realistic depictions of individuals saying and doing things that never genuinely occurred. The process involves aligning the faces of two different people, using an autoencoder to capture characteristics from one face (identified as "face A"), and subsequently merging these characteristics with another face (identified as "face B"). This results in the creation of a face that looks similar to B but does not authentically depict their actual appearance (Alanazi & Asif 2023). Such facial reconstruction techniques are exploited in illicit activities, particularly for creating adult or explicit content on the black market. Deepfakes rely on neural networks that analyze extensive datasets to acquire the ability to mimic human facial features, expressions and voice, making it exceedingly difficult for people to differentiate between real and fake content. Furthermore, producing convincing fake content does not necessarily require expertise, as non-experts can create such deep fakes using readily available tools like Face2Face and FaceSwap.

Regrettably, deepfakes are frequently utilized for malicious purposes, including scams, such as impersonating the voices of business professionals or deploying them in reputation-damaging situations, like politics and deceptive contexts.

Given these challenges, it is vital to explore the utilization of detection techniques and effective methods to mitigate the possible hazards associated with deepfake technology. The aim of this review paper is to conduct comprehensive research on the production and identification of deepfakes to gain a better understanding of this technology. In doing this, it aims to clarify the complex aspects of this mysterious and worrisome technology, providing valuable insights for navigating it and protecting against its possible negative outcomes.

The first part of this paper explores the generation of deepfakes within the realm of deepfake technology.

✉ Sami Alanazi
sami.alanazi@cranfield.ac.uk

Seemal Asif
s.asif@cranfield.ac.uk

[1] Cranfield School of Aerospace Transport and Manufacturing, Cranfield University, Wharley End, UK

Published online: 18 September 2024

Subsequently, the variety of available software and apps behind deepfake creation is investigated. Following that, deepfake detection is discussed, consisting of two parts: fake image detection and fake video detection. The fourth section of this paper focuses on the manipulation of images and videos that involve human expressions within the realm of deepfakes. Afterward, the social impact and legislation surrounding deepfakes are examined. Finally, the paper concludes with a summary of key findings and insights.

2 Generating deepfake

Deepfakes are produced using deep neural networks, specifically through the utilization of autoencoders (Juefei-Xu et al. 2022). This procedure involves the training of a neural network to encode and decode images or videos, as depicted in Fig. 1. The encoder's role is to take the initial input of an image or video and condense it into a latent code, retaining the critical features while filtering out unnecessary details. Subsequently, this latent code is transmitted to the decoder, which reconstitutes the original content based on this code (Nguyen et al. 2019).

In the process of generating fabricated content, the autoencoder is trained with both authentic and altered videos or images. The encoder learns to encode both real and deepfake materials, producing comparable latent representations for each type. Simultaneously, the decoder uses these forged latent codes to reconstruct the initial input, ultimately facilitating the production of highly convincing deepfake content. The generation of such deepfake content relies on a range of technologies, including algorithms like 3D ResNeXt and 3D ResNet (Alanazi & Asif 2023).

Generative adversarial networks (GANs) represent a powerful class of deep neural networks increasingly utilized for creating deepfake content, such as counterfeit images and videos (Malik et al., 2022). A typical GAN architecture includes two main components: a generator and a discriminator. The generator crafts new data samples, whereas the discriminator evaluates them against real data to distinguish authenticity. Throughout the training process, the generator strives to fool the discriminator, which in turn adapts to better identify fake data. This interplay, however, faces limitations when working with small datasets, requiring substantial data volumes to function effectively and reliably, as noted by Almars (2021).

The prevalence of altered images and videos underscores the importance of reliable detection techniques for distinguishing between genuine and counterfeit content. In this regard, Yang and colleagues (2022) propose a method known as deepfake network architecture attribution, which identifies the specific generator architectures behind the creation of counterfeit images. This approach remains effective even when used on advanced models that have undergone retraining across multiple datasets.

Delving deeper into deepfake technology, especially the attribution of network architectures as shown in Fig. 2, attribution can be approached at two distinct levels: the architectural and the model specific. This study assesses two methodologies: one leveraging learned features and

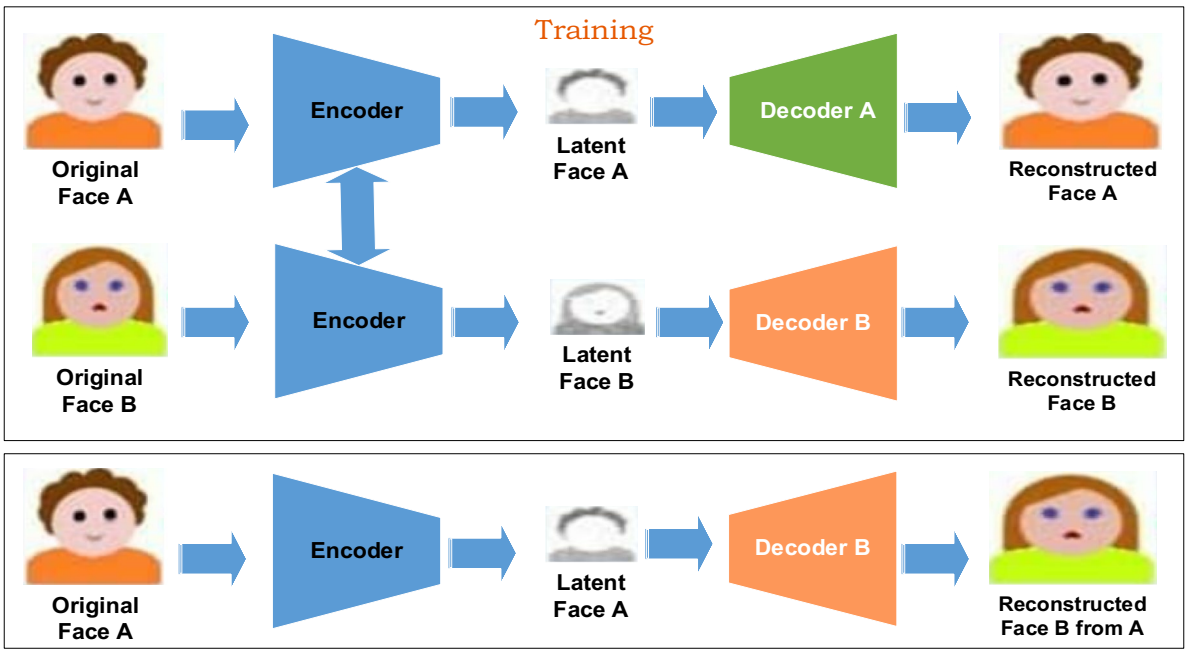

Fig. 1 Deepfake model: autoencoders with dual pairs

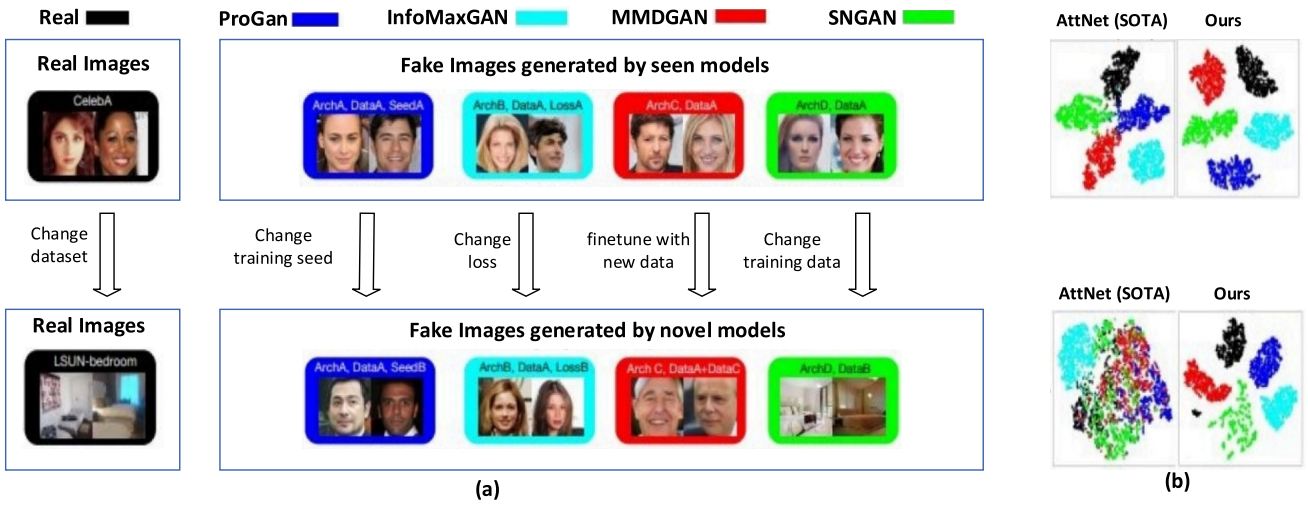

Fig. 2 Deepfake generation (Yang et al. (2022))

another utilizing AttNet. AttNet isolates unique attributes from GAN-generated images, showing distinct effectiveness when comparing generated and real images from consistent GAN models and training sets. However, AttNet's effectiveness diminishes with novel or modified training scenarios, unlike the proposed method which maintains its discriminative power, as detailed in studies by Yu et al. (2018) and further analyzed in Yang et al. (2022), with the differences in feature extraction capabilities visually represented through t-SNE analysis.

The deepfake content generation generally follows this principle where the deepfake images and videos are relatively less clear as compared to the real images and videos used to create the output. The fabricated content is less in resolution, but the lay man in first impression can mistake it as the real content. Deepfake combines features from different sources to create an output that looks like the real one but has few major changes which alter the meaning of the overall image or the video. That feature can be smile, cry, body part, expression or colour of skin.

3 Tools and software for creating deepfake content

The swift progress in deepfake creation applications, fuelled by its demand in underground markets underscores the need for ongoing enhancements in detection techniques (Shahzad et al. 2022). Numerous tools are now accessible for producing deepfake content, and a selection of them is provided below.

A well-known tool, DeepSwap, is favoured for generating fabricated content for recreational purposes. It is known for its user-friendly nature and easy online accessibility. Many users prefer the free version, which can be installed on both mobile devices and laptops. This tool is notable for offering two key features. Firstly, it operates with remarkable speed, making it possible to generate realistic-looking content in a notably brief timeframe (Wilpert 2022). Its efficiency ensures quick results. Moreover, the images it produces closely mimic genuine ones, making it difficult for viewers to initially distinguish between authentic and counterfeit content (Rankred 2022).

DeepSwap strictly enforces its terms of service, explicitly prohibiting the creation or sharing of pornographic deepfakes. It mandates that users must not upload, share or transmit any inappropriate content (De Silva De Alwis & Careylaw, 2023). Despite its capabilities, the application has faced criticism from users who find it challenging to unsubscribe, as the process for terminating subscriptions is perceived as overly complicated. This has led to only a limited number of users recommending DeepSwap within their social circles; there have been user complaints about difficulties in unsubscribing from the application as it appears to make the termination of subscriptions complicated. Consequently, only a small subset of users tends to endorse the tool to individuals within their social networks.

DeepFace Lab is a platform frequently utilized by students and researchers to produce altered images and videos on computer systems. While it might not be as approachable for the general public, it is highly appreciated by researchers for its adaptability in selecting the machine learning technology employed (Wilpert 2022). The interface is uncomplicated, although it holds particular value for researchers with programming proficiency. Furthermore, the software is compatible with computers featuring diverse processing capacities, expanding its accessibility to a broader range of users (Rankred 2022).

DeepFace Lab excels in producing remarkably realistic outputs and serves as an open-source tool for realistic face swapping, including advanced capabilities like de-aging faces in images. While invaluable to researchers, models and actresses, its complex interface may be less user-friendly for non-technical users.

DeepFace Lab initially employed a subject-aware encoder-decoder method for face swapping that was restricted to two specific identities (Xu et al. 2022). However, more recent developments have introduced subject-agnostic approaches that simplify the process and increase its versatility (Xu et al. 2022). These methods are divided into two categories: source-oriented, focusing on the characteristics of the original video, and target-oriented, adapting to the features of the destination video. This state-of-the-art technology, coupled with DeepFace Lab's integrated and user-centric design as described by Perov et al. in 2020, not only simplifies the creation of photo-realistic face-swapping videos but also supports diverse computational setups. Its scalability, efficient resource utilization and broad adaptability enhance both creative video production and digital forensics, establishing it as a crucial tool in both entertainment and technological fields.

DeepNostalgia is a popular deepfake application known for its capability to produce high-resolution images and videos that mimic authentic visuals with impressive accuracy. Its clear image quality and photo enhancement functionality make it particularly attractive to users interested in crafting engaging content and sharing emotionally animated portrayals. As noted by Kidd and Nieto McAvoy (2023), this technology not only enhances the quality of vintage photographs but also brings them to life by animating them with realistic gestures based on actual human actions. Although DeepNostalgia is popular for its user-friendly features that facilitate easy sharing across social networks, it has also sparked debate over ethical issues, particularly the animation of deceased individuals and the potential for commercial misuse (Kidd & Nieto McAvoy 2023). This intricate interplay between technological advancement and ethical considerations underscores the profound influence that digital tools have on personal and collective memory, prompting a deeper investigation into their implications in modern genealogy and social dynamics.

Deep Art Effects is accessible for both computers and mobile platforms, although mobile users tend to express dissatisfaction with the results. Compatibility problems, including issues with iPhones, have been noted. Although the commercial version is considered more effective, the free version is not well-received. Its limited popularity as a deepfake tool is further exacerbated by problems with refunds and inconvenient image selection (Wilpert 2022). Table 1 offers a detailed comparison of each discussed tool,

Table 1 Comparison of deepfake tools by features, user base and limitations

Tool	Target users	Platform compatibility	Accuracy/ease of use	Key features	Limitations
DeepSwap	General public	Mobile and laptops	Very user-friendly	Fast processing, realistic output, enforces terms against inappropriate content	Complicated subscription cancellation, limited endorsements
DeepFaceLab	Researchers, models, actresses	Computer systems	Complex for non-technical users	High-quality output, face-swapping, de-aging, adaptable technology, subject-aware and subject-agnostic methods for increased versatility	Requires programming proficiency, not intuitive for the general public
DeepNostalgia	General public, genealogy enthusiasts	Primarily web-based	User-friendly	High-resolution, photo enhancement, animates vintage photos with realistic gestures	Ethical concerns, especially with animating deceased individuals
DeepArtEffects	General public	Computers and mobile	Challenging on mobile	Available on multiple platforms, the commercial version is more effective	Compatibility issues, dissatisfaction with the free version, refund issues

elucidating their respective capabilities, features and potential limitations.

4 Deepfake detection

The growing threat posed by deepfakes to privacy, security and democracy. In response to this emerging danger, various methods have been proposed to detect deepfakes. Initial efforts relied on spotting artificial traits stemming from glitches and inconsistencies in artificially created videos. In contrast, more recent methods have harnessed deep learning to extract meaningful and distinguishing traits to identify deepfakes (Chesney and Citron 2019).

Typically, the problem of detecting deepfakes is approached as a binary classification task, where the goal is to differentiate between genuine and fabricated videos. However, this procedure necessitates a substantial dataset of both real and forged videos to train classification models (de Lima et al., 2020). Although counterfeit videos are becoming increasingly prevalent, there is a notable absence of established benchmarks for evaluating a range of detection methods. In an effort to address this issue, Korshunov and Marcel (2018) have created a noteworthy dataset specifically designed for evaluating deepfakes. This dataset comprises 620 video models generated using the open-source FaceSwap-GAN code. To create this dataset, publicly accessible films from the VidTIMIT database were employed. These films were utilized to generate deepfake videos characterized by realistic facial expressions, mouth movements and eye blinks. Subsequently, these videos served as the basis for evaluating a variety of detection techniques.

The test results indicate that well-known facial recognition systems relying on VGG and Facenet face difficulties in accurately detecting deepfakes. Furthermore, techniques like lip-sync analysis and image quality assessments utilizing support vector machines (SVMs) manifest a notably elevated error rate when employed for the identification of deepfake videos within this freshly generated dataset. These findings underscore the pressing need for the development of more robust approaches for deepfake detection (Wen, Han, and Jain 2015). Subsequent sections will outline different categories of deepfake detection methodologies.

5 Fake image detection

Face-swapping technology offers numerous practical applications in video editing, portraiture and safeguarding privacy by allowing the replacement of faces in images with others from a photo collection. However, it has also been exploited by cybercriminals for unauthorized access and identity theft (Korshunova et al. 2017). Modern deep learning techniques, such as convolutional neural networks (CNNs) and generative adversarial networks (GANs), have made it challenging to detect swapped facial images because they can retain facial features like position, expressions and lighting. To address this issue and differentiate between authentic and altered facial images, Zhang et al. (2017) employed a method referred to as the "bag-of-words" technique to extract compact features, which were then input into various classifiers like support vector machines (SVMs) and multi-layer perceptrons (MLPs). Among various types of manipulated images, GAN-generated deepfakes pose a particularly tough challenge owing to their exceptional quality, realism and the GAN's capacity to simulate intricate data distributions and produce outcomes that closely resemble the input data distribution.

Regarding the detection of GAN-generated deepfakes, Agarwal and Varshney (2019) approached it as a hypothesis testing problem, considering it a statistical framework rooted in information theory and authentication research. They determined the "oracle error", which is the minimum distance between the distribution of genuine images and images produced by a specific GAN. Their analysis revealed that as the GAN's accuracy diminishes, this distance expands, facilitating the detection of substantial imperfections in deepfakes. This is particularly pertinent when dealing with high-resolution image inputs, where GANs are crucial in crafting fraudulent images that are exceedingly challenging to distinguish (Nguyen et al. 2019).

6 Fake video detection

Detecting fake videos poses unique challenges due to the degradation of frame data during video compression and the temporal characteristics inherent to videos. Many traditional image identification methods are ill-suited for video analysis, primarily because videos exhibit temporal characteristics that go beyond still frames. This makes it necessary to develop techniques specifically tailored for detecting video deepfakes (Afchar et al. 2018).

One approach to deepfake video detection involves analyzing the temporal properties of video frames. Sabir et al. (2019) leveraged the spatio-temporal characteristics of video streams to uncover inconsistencies introduced during the deepfake synthesis process. They performed frame-by-frame analysis to reveal low-level anomalies caused by facial alterations, which manifest as temporal contradictions between frames. Their method comprises two main steps: initially, identifying, cropping and aligning faces within a sequence of frames, and subsequently, employing a fusion of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to differentiate between manipulated and genuine facial images, as illustrated in Fig. 3 by Nguyen

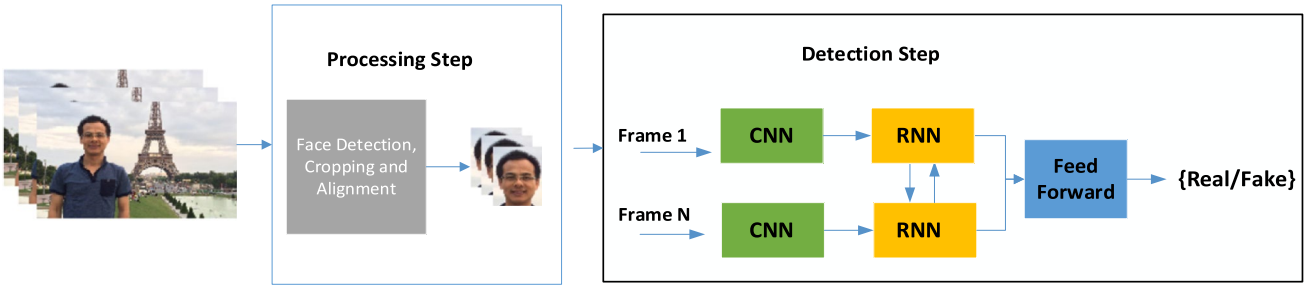

Fig. 3 The process for detecting facial manipulation in two sequential steps (Nguyen et al. 2019)

et al. (2019). This approach was evaluated on the FaceForensics++dataset, consisting of 1000 videos, yielding promising results.

Another method for fake video detection focuses on analyzing individual video frames to identify visual characteristics that can differentiate real videos from deepfake ones. Afchar et al. (2018) introduced Meso-4, a deep learning technique that utilizes a complex architecture involving convolutional and pooling layers to identify elements of deepfake content. MesoInception-4 is an improved version of Meso-4, incorporating the inception module to enhance model performance. While Meso-4 excels in binary classification and distinguishing between deepfake and authentic images, it is built on a relatively shallow CNN architecture, potentially limiting its capacity to identify intricate manipulations. Neural networks have proven effective in deepfake detection, with an emphasis on identifying artifacts related to facial warping and physiological/biological features. Ciftci and Demir (2020) discuss that the central focus of deep fake content as well as surrounding areas help in detection of deep fake. One approach is the detection of Face Warping Artifacts which involves analysing processed face areas and neighbouring content to observe deepfake algorithms and generate images of limited resolutions that can be used to match fake content with the source content (Jadhav et al. 2020). The key feature in the creation of deep fake is copy-pasting of selected features from the original content into the processed and fake content. On the other hand, the solution lies in noise detection and finding how certain content might differ from the original. The creators of deep fake content focus on prominent features of a face such as eyes, lips and nose, but the detection of deepfakes requires the use of more complex and unique features of a person, such as eye blinking. Hence, (Jadhav et al. (2020) discusses that exposing deepfake requires utilizing physiological and biological features that go beyond the observations of criminals behind fake content. In their recent work (Raza et al. 2022), they presented a deepfake detection model. This model was trained on a dataset comprising both counterfeit and authentic human faces, resulting in a notably high level of accuracy in the detection of deepfake elements.

The availability of deepfake datasets, often sourced from platforms such as Kaggle (n.d.), has facilitated the training and evaluating of neural network techniques for deepfake detection. These models employ transfer learning, making use of pre-trained models to discern between authentic and manipulated images by scrutinizing facial characteristics. Algorithms scrutinize various aspects such as dimensions, size and shapes of facial features to spot inconsistencies and categorize images or videos as forged.

One specific approach, the Xception Technique, relies on transfer learning-based neural networks and employs deep separable convolution layers to identify changes in both images and videos. The efficacy of different deepfake detection methods may fluctuate depending on factors such as the dataset's size and the complexity of the algorithm. Promising approaches include pro-3D CNN and physiological measurements, such as heart rate assessment using long-distance photoplethysmography (rPPG), although they require further development. Researchers are also exploring meta-learning techniques for deepfake detection.

It is important to acknowledge that the current forensic processes are often complex and time-consuming. Therefore, there is an increasing demand for more streamlined tools that can confirm the legitimacy of videos and images. Deep learning techniques have substantial potential in discerning between counterfeit and authentic content, but further progress is essential to tackle the issues presented by deepfake technology.

Distributed ledger technologies (DLTs) dig into and identify the origin of a video which helps in contributing to preventing deepfake content. When the basic roots or features of a video are identified, the real video can be identified among fake videos. Deepfake videos are made by tampering with certain elements in a video, not all of the features in a video. This leaves room for the identification of deepfakes (Zichichi et al. 2022). In DLTs, every transaction is assigned certain order in a way that every participant can use those transactions in exact order to a certain shared state hence guaranteeing that all the copies of the state remain consistent.

7 Altering images and videos with human emotions in deepfake content

Modifying static images is generally simpler than working with moving images. Nonetheless, manipulating videos featuring human expressions presents a notable challenge in the realm of deepfake content manipulation. Every person has their distinct manner of expressing themselves, and when combined with their facial characteristics, it leads to unique visual results. Deepfake videos, as described by Groh et al. in 2021, are typically created from publicly available datasets where human faces often appear devoid of any meaningful expressions, resembling lifeless puppets. To overcome this constraint, advanced deepfake technologies have arisen, emphasizing the alteration of a wide range of motions, including facial and bodily gestures and expressions. Machine learning is utilized to simulate human actions such as walking, speaking, grinning, sobbing and scowling. These models are then used to replace the original identity. It is important to note that manipulating videos with fewer expressions and shorter durations is simpler compared to those featuring complex expressions, multiple variations and longer durations.

Advanced algorithms incorporate principles from psychology, probability, kinematics, inverse kinematics and physics to identify deepfake content by scrutinizing the temporal elements of videos. In the domain of deepfake detection, neural network algorithms that prioritize facial localization, such as CNNs, have demonstrated remarkable accuracy. Their focus lies in facial positioning rather than consistent emotional speech and expressions, as discussed by Groh et al. in 2021.

The process of identifying deepfake manipulation involves a thorough examination of specific facial regions rather than the whole image. Algorithms utilize fusion methods to spot alterations by contrasting these regions with an extensive training dataset that covers facial traits across diverse demographics. A variety of attributes, such as facial expression, hair and eyes, are employed as random markers to assess changes. Even subtle distortions in facial regions, which may go unnoticed by humans, can significantly impact the final image. Algorithms are dedicated to closely monitoring these selected regions for precise detection, as highlighted in the works of Tolosana et al. in 2022 and Guarnera et al. in 2022.

Detecting deepfake content involves more than just focusing on the depicted individual, it also includes considering background and scene elements. Algorithms are designed to recognize changes within scenes, beginning with straightforward backgrounds and progressively addressing more complex scenarios. Scene element rotations and insights from domain experts contribute to the recognition of crucial attributes unique to particular contexts. Detecting alterations in these features allows algorithms to categorize deepfake images by the identified changes, as described by Choras et al. in 2020 and Siegel et al. in 2021.

Data scientists and artificial intelligence experts are actively researching techniques for identifying counterfeit images and videos by scrutinizing both conspicuous characteristics like accents and subtle aspects like lighting conditions. Training datasets are meticulously designed to highlight elements such as poses, postures, lighting conditions and backgrounds to evaluate authenticity. The inherent principles of lighting physics offer promising prospects for detecting deepfakes, even though artificial intelligence tools are still evolving in this domain. Ongoing research is dedicated to improving deepfake forensics by delving into the physics of lighting (Somers 2020).

Nirkin et al. 2022) discuss that face swapping can lead to manipulation of face region that leads to adjusting a face in a new context. The same method can be used to keep the scenario and background while swapping the face only. In either case, the person whose face is used will be shown to be a part of an event that he was not a part of. The detection of this type of manipulation can be done by carefully observing certain indicative signs of manipulation. The face's context of hair, ears, neck, etc. can be monitored to detect copy-paste or other manipulations. Liu et al. (2021) discuss that the consistency of the image changes when it is manipulated; hence, face swap also results in certain inconsistencies that can be detected using the face swap method. Liu et al. (2021) argue that a forensic specialists must know inconsistencies that result from face swapping because only then they will be in a position to look for the right clues that lead to deepfake detection. This involves fine grain abnormalities in areas/boundaries where face-swapping is suspected.

The advancement of generative adversarial networks (GANs) has raised significant apprehensions regarding the privacy and trust of online users, mainly because of their capability to produce exceptionally convincing deepfake content. GANs improve manipulated images by incorporating adversarial and perceptual losses, yielding visually persuasive forgeries. Techniques like frame-to-frame face detection and facial reenactment contribute to the heightened realism of videos produced through GAN processing. Among common deepfake methods, face morphing and face swapping are notable, with face morphing involving the fusion of features from multiple individuals. Detecting morphed facial images is essential for reliable recognition systems, and methods like morphing attack detection (MAD) can be employed. GANs play a substantial role in the creation of counterfeit data and the manipulation of images, producing high-resolution fake images that are difficult to discern from genuine ones. Techniques like deep convolution

generative adversarial networks (DCGAN) are valuable for the training of GANs to generate more convincingly deceptive images.

To detect deepfake videos, phoneme-viseme mismatches are used, where the spoken sound does not align with the mouth's shape (Agarwal et al. 2020). These subtle yet significant inconsistencies are helpful in spotting manipulations, and language specialists are frequently consulted to detect deepfakes in various languages. Forensic methods that rely on human expertise are employed, with the support of deep learning algorithms to aid in the decision-making process. Attention-based explainable deepfake detection algorithms enable experts to concentrate their attention on specific regions within images and videos. The human intuition and consideration of cultural context are additional elements contributing to the detection of deepfakes. Forensic experts take a hands-on approach by manually selecting specific regions within content, which can subsequently undergo further processing using software tools to enhance the accuracy of detection.

Forensic technique for the detection of deep fake is used where human involvement is required. Silva et al. (2022) discuss that forensics algorithms depend upon human effort who use deep learning detection algorithm and help in making decisions regarding whether the content is original or fake. There are several forensic techniques, and Silva et al. (2022) are in favour of an attention-based explainable deepfake detection algorithm which helps in deploying detection networks to detect faces and other elements of images and videos. Humans can choose which region to ignore, enlarge or focus more while detecting deepfake content. There are several aspects of images and videos which can be assessed in certain pretexts. People understand their cultural and social pretexts better than machines in many cases. Hence, human involvement and forensic technique are commonly used to detect deepfake. Human instincts also play a role in this technique of detection. The regions that are manually selected by the forensic experts can then be processed using tools and software so that deepfake can be accurately detected finally.

Face morphing and face swap are two main techniques used in deepfake to alter images or video in order to produce counterfeit content. The key difference between them is face swap process involved replacing the face of one person in an image or video with someone else face, while the face morphing procedure involves blending the facial features of more than two people to create a new hybrid face. Face morphing is a challenge for recognition systems; hence, it is critical to develop methods for identification of facial morphing.

The danger of face morphing technique in deep fake technologies lies on its malicious use. This can be done by morphing a real image of themselves and a companion and blending the facial features to produce morphed image as their photograph for an ePassport (Dameron 2021). This allows them to appear as the accomplice and pass through the checkpoint without raising any red flags, even if they are wanted by the authorities (Dameron 2021). Therefore, it is critical to detect fake images created using this technique. Damer et al. (2019) proposed a detection method called landmark-based solution by utilizing the live probe image of a potential attacker's face as an additional source of information. The authors' concept targets the facial landmarks in both the reference and live probe images. The proposed solution assumes that it is possible to recognize specific patterns in the changes of facial landmarks' position in the two images when a morphed reference is used. Damer et al. (2019) explain the workflow of the landmarks-based solution approach as illustrated in Fig. 4.

The process starts with scanning the facial landmarks in both the reference and probe image to create a features vector based on the shifts in the landmark's location. This vector is then used to classify the reference image as either a morphing attack or a bona fide image. Damer et al. (2019) present examples of landmarks shifts in attack and bona fide image pairs, along with a description of the techniques employed for facial landmark detection. Figure 5 shows these examples by Damer et al. (2019) of the facial landmarks in bona fide and reference images of the same two subjects, along with their corresponding probe images.

8 Deepfake social impact and legislation

Deepfake videos were originally considered a form of amusement, anticipated to be enjoyed by both those who made them and those who appeared in them. Moreover, film production companies are starting to widely utilize deepfake technology to edit scenes, which allows them to avoid the expenses and time associated with reshooting (Uddin Mahmud & Sharmin 2020).

Nevertheless, deepfake technology quickly began to be used for creating explicit material and potential blackmail, raising significant social concerns. According to a study by Hancock and Bailenson (2021), one of the major negative effects of this technology: the undermining of public trust in media. Such videos and images also promote manipulation and deceit, leading to widespread uncertainty about the authenticity of visual evidence. Deepfakes can distort personal memories and implant entirely false ones, potentially changing one's perception of others without any real basis (Hancock and Bailenson 2021).

As technology continues to advance, new methods of committing crimes are also emerging. Current laws frequently prove inadequate for addressing the challenges posed by these novel forms of criminality, underscoring

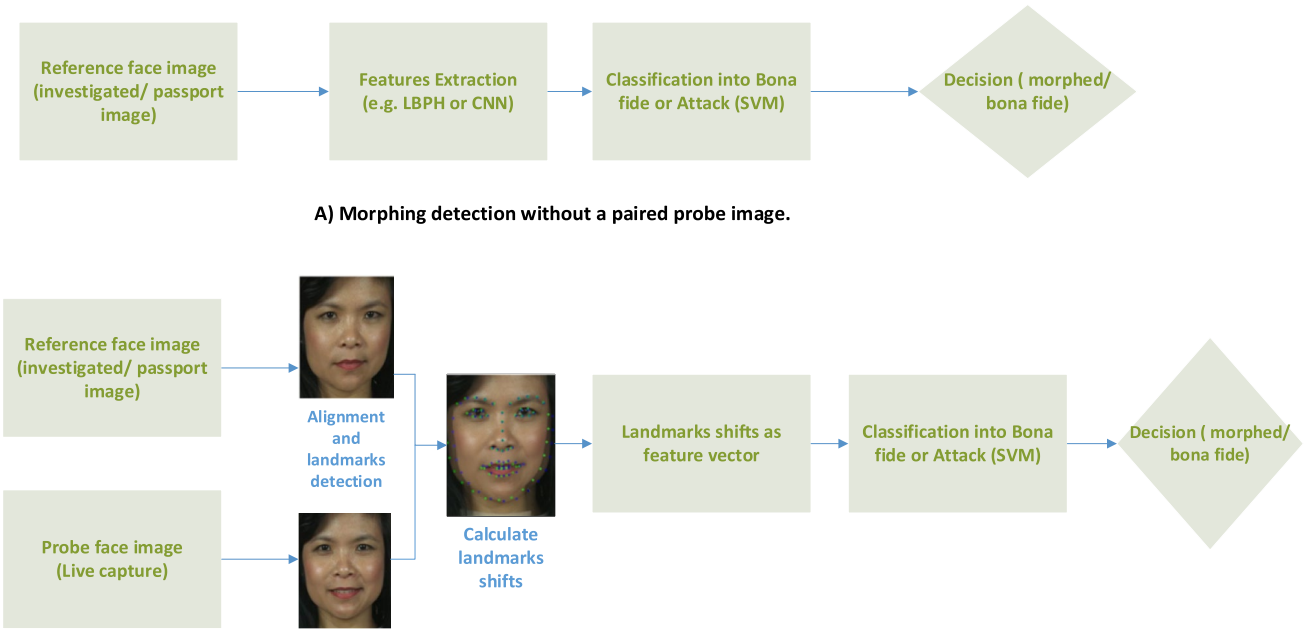

Fig. 4 Landmarks-based solution workflow (Damer et al. 2019) (**A**, **B**)

Fig. 5 Morphing attacks detection (Damer et al. 2019)

the necessity for updated and more sophisticated legislation that comprehensively addresses cybercrimes and imposes appropriate penalties on wrongdoers. The damage potential of deepfakes became starkly evident in situations like the 2018 Rohingya genocide in Myanmar, believed to be fueled by deepfake-generated content (GOV.UK 2019). During Kenya's 2018 elections, there was speculation that deepfake videos of an unwell presidential candidate were spread to influence public perception falsely (Kigwiru 2022; van der Sloot and Wagensveld 2022).

The UK government has recognized the need for specific regulations targeting various forms of deepfakes, including face reenactment, face generation and speech synthesis (GOV.UK 2019). With the growing complexity of deepfake technology, identifying and penalizing such content present greater difficulties. Legislation is being

formulated to deter the creation of deepfake content for political and societal manipulation recognizing that it has the potential to inflict harm and impact the standing and livelihoods of individuals, entities and political groups (GOV.UK 2019). Additionally, the European Union's AI Act is part of a broader effort to enforce transparency and ensure that users are fully informed when interacting with AI systems capable of creating or modifying media content such as deepfakes. The Act stipulates varying requirements based on the risk associated with the AI system involved, aiming to protect users and enhance their ability to make informed decisions (Europarl 2023). The EU's legislative approach, encapsulated by the Artificial Intelligence Act, continues to stress the importance of transparency and the protection of fundamental rights to prevent risks associated with AI, such as media manipulation (EC 2024; Loughran 2024).

American courts are increasingly recognizing the threat posed by deepfake content in criminal activities. This has led various states to enact specific legislation targeting the misuse of this technology. In Texas, for example, amendments made in 2019 to Sect. 255.004 of the Election Code now regulate the production and distribution of deepfake videos during state elections (Kigwiru 2022). Violations of this law carry severe consequences, including up to one year in county jail and fines of $4000, underscoring the gravity with which Texas treats the potential election-related abuses of deepfake content (Kigwiru 2022).

These state-level legislative efforts are part of a broader pattern of regulations across different regions aimed at combating the misuse of AI technologies and addressing deceptive practices. In the USA, the Federal Communications Commission (FCC) has banned AI-generated robocalls that impersonate public figures, which is part of a larger initiative against digital fraud (Kan 2024; Yousif 2024). Similarly, in China, the Cyberspace Administration has enacted regulations that prohibit the unauthorized creation of deepfakes. These laws also mandate that AI-generated content be clearly labelled, a measure that helps protect personal privacy and national security (CAC 2022).

Given these factors, there is a need for thorough legislation that targets the production of deepfake content and penalizes offenders not only for their actions but also for the damage inflicted on the victims. Such harm may encompass psychological distress damage to one's reputation or even electoral losses resulting from the dissemination of misinformation via deepfakes. Additionally, media outlets and government bodies should initiate educational campaigns to foster a more discerning and informed society, protecting it from the disruptive influence of a deepfake (Alanazi et al. 2024).

9 Discussion and conclusion

The rapid advancement of deepfake technology has raised worries about its potential for deceit and unethical applications. In order to protect online users and ensure a secure digital space, legislative measures are being put in place. While identifying deepfake content remains challenging, researchers have discovered indicators that can assist in this process, such as abnormal eye blinking patterns. Realistic blinking was initially absent from deepfake systems, but more recent methods have incorporated it. The process of identifying deepfakes can be intricate, involving the training of machines to differentiate between various blink patterns for different individuals and situations. The detection and prevention of deepfake content are being enhanced through the utilization of artificial intelligence (AI) and other advanced technologies. Even when the disparities in appearance between genuine and fake content are subtle, machine learning algorithms have the capability to discern anomalies in facial expressions and eye blinking. These advancements underscore the importance of employing technology for deepfake detection rather than relying solely on human observation.

Deepfake technology offers both benefits and drawbacks. Policymaking is imperative to mitigate the risks associated with deepfake content, encompassing state-level regulations, policies on social media platforms and national laws to penalize those who produce and distribute deepfakes with malicious intent. Public awareness campaigns are essential to educate the public about the ethical boundaries related to deepfake content. Effective collaboration among governments, technology companies and the public is necessary to develop methods for detecting and preventing deepfakes. The field of cyber law enforcement should adapt to ensure the security of all online users. Sustained innovation and the implementation of regulatory measures are essential to tackle the issues posed by deepfakes. As depicted in Fig. 6, the workflow of deepfake content comprises its generation, dissemination across social media platforms, the detection process and the execution of measures to monitor it. This process involves policymaking, awareness campaigns and the collaboration mentioned above. It is essential to underscore the significance of establishing a feedback loop between detection and mitigation to effectively monitor the spread of deepfakes.

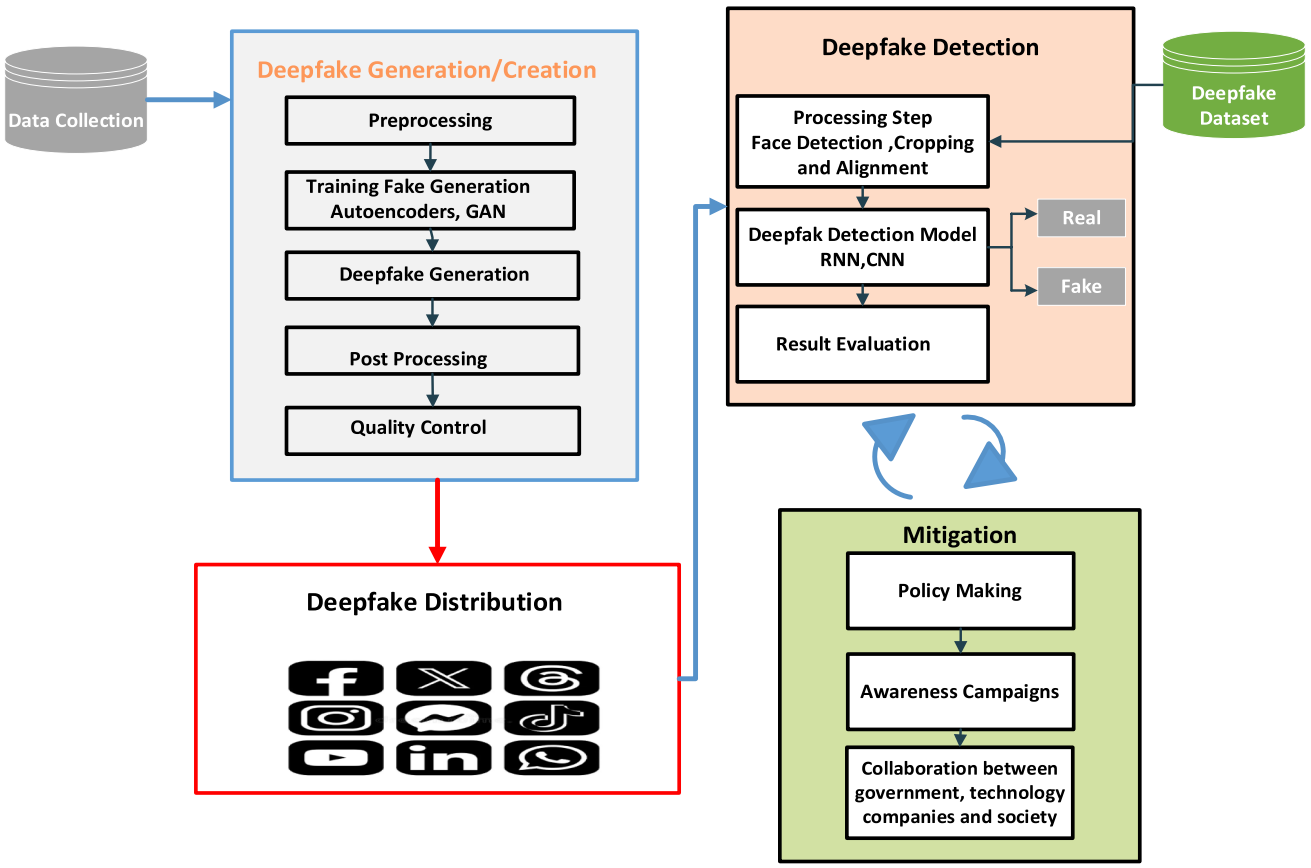

Fig. 6 Deepfake lifecycle diagram

Declarations

Conflict of interest The authors declare no competing interests.

References

Afchar, D. et al. (2018) 'MesoNet: a compact facial video forgery detection network'. Available athttps://doi.org/10.1109/WIFS.2018.8630761

Agarwal, S. and Varshney, L. R. (2019) 'Limits of deepfake detection: a robust estimation viewpoint'. Available at: https://arxiv.org/abs/1905.03493

Agarwal, S. et al. (2020) Detecting deep-fake videos from phoneme-viseme mismatches. Available at: https://ieeexplore.ieee.org/document/9151013

Alanazi, S., and S. Asif. 2023 'Understanding deepfakes: a comprehensive analysis of creation, generation, and detection'. Available at: https://doi.org/10.54941/ahfe1003290.

Alanazi S, Asif S, Moulitsas I (2024) Examining the societal impact and legislative requirements of deepfake technology: a comprehensive study. IJSSH. https://doi.org/10.18178/ijssh.2024.14.2.1194

Almars, AM (2021) 'Deepfakes detection techniques using deep learning: a survey', J Comput Commun, 09(05), pp. 20–35. Available at: https://doi.org/10.4236/jcc.2021.95003.

De Silva De Alwis R, Careylaw P (2023) A rapidly shifting landscape: why digitized violence is the newest category of gender-based violence. In Public Law and Legal Theory Research Paper Series Research Paper (Issue 23). https://ssrn.com/abstract=4648409Electroniccopyavailableat, https://ssrn.com/abstract=4648409

CAC. (2022, December 12). The Cyberspace Administration of China and Other Three Departments Issued the "Provisions on the Administration of Deep Synthesis of Internet Information Services." Cyberspace Administration of China. http://www.cac.gov.cn/2022-12/11/c_1672221949318230.htm

Chesney B, Citron D (2019) Deep fakes: a looming challenge for privacy, democracy, and national security. Calif L Rev 107(6):1753–1820. https://doi.org/10.15779/Z38RV0D15J

Choras, M. et al. (2020) 'Advanced machine learning techniques for fake news (online disinformation) detection: a systematic mapping

study', Applied Soft Computing [Preprint]. Available at: https://arxiv.org/abs/2101.01142
Ciftci, UA, Demir I (2020) 'FakeCatcher: detection of synthetic portrait videos using biological signalshttps://doi.org/10.1109/TPAMI.2020.3009287
Damer, N et al (2019) 'Detecting face morphing attacks by analyzing the directed distances of facial landmarks shifts', in lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics). Springer Verlag, pp. 518–534. https://doi.org/10.1007/978-3-030-12939-2_36
Dameron, J.L. (2021) Real vs fake faces: DeepFakes and face morphing. Available at: https://researchrepository.wvu.edu/etd/8059.
EC. (2024). AI Act. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai
Europarl, (2023) EU AI Act: first regulation on artificial intelligence, Europarl, Available at: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence (Accessed: 15 September 2023).
GOV.UK (2019) Snapshot paper - deepfakes and audiovisual disinformation, Available at: https://www.gov.uk/government/publications/cdei-publishes-its-first-series-of-three-snapshot-papers-ethical-issues-in-ai/snapshot-paper-deepfakes-and-audiovisual-disinformation (Accessed: 9 February 2023)
Groh M et al (2021) 'Deepfake detection by human crowds, machines, and machine-informed crowds', arxiv [Preprint]. Available at: 10.1073/pnas.2110013119.
Guarnera, L. et al. (2022) 'The face deepfake detection challenge', Journal of Imaging, 8(10)https://doi.org/10.3390/jimaging8100263
Hancock JT, Bailenson JN (2021) 'The social impact of deepfakes', cyberpsychology, behavior, and social networking. Mary Ann Liebert Inc., pp. 149–152. Available at: https://doi.org/10.1089/cyber.2021.29208.jth
Jadhav A et al (2020) Deepfake video detection using neural networks, IJSRD-International Journal for Scientific Research & Developmentl. Available at: www.ijsrd.com
Kaggle. (n.d.). Kaggle. Retrieved January 18, 2024, from https://www.kaggle.com
Kan M (2024). Biden Calls for a Ban on AI Voice Impersonation. https://uk.pcmag.com/ai/151364/biden-calls-for-a-ban-on-ai-voice-impersonation
Kidd J, Nieto McAvoy E (2023) Deep nostalgia: remediated memory, algorithmic nostalgia and technological ambivalence. Convergence 29(3):620–640. https://doi.org/10.1177/13548565221149839
Kigwiru VK (2022) Deepfake technology and elections in Kenya: Can legislation combat the harm posed by Deepfakes?. Available at SSRN: https://doi.org/10.2139/ssrn.4229272 or https://ssrn.com/abstract=4229272
Korshunov, P. and Marcel, S. (2018) 'DeepFakes: a new threat to face recognition? Assessment and detection'. Available at: https://arxiv.org/abs/1812.08685
Korshunova I et al (2017) 'Fast face-swap using convolutional neural networks'. Available at: https://arxiv.org/abs/1611.09577
De Lima O et al (2020) 'Deepfake detection using spatiotemporal convolutional networks'. Available at: https://arxiv.org/abs/2006.14749
Liu K et al (2021) 'Face swapping consistency transfer with neural identity carrier', Future Internet, 13(11)https://doi.org/10.3390/fi13110298
Loughran J (2024) EU signs law to crack down on 'high risk' AI. Eandt. https://eandt.theiet.org/2024/03/14/eu-signs-legislation-crack-down-high-risk-ai?utm_campaign=E%2BT%20News%20-%20Template%20Redesign%2014%20Mar%20%28Split%20test%29&utm_content=E%26T%20News%20-%20Members&utm_medium=email&utm_source=Adestra&utm_term=3477864
Malik A, Kuribayashi M, Abdullahi SM, Khan AN (2022) DeepFake detection for human face images and videos: A survey. In IEEE Access (Vol. 10, pp. 18757–18775). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2022.3151186
Nguyen, ThanhThi et al (2019) 'Deep learning for deepfakes creation and detection: a survey'https://doi.org/10.1016/j.cviu.2022.103525
Nirkin Y, Keller Y, Hassner T (2022) 'FSGANv2: improved subject agnostic face swapping and reenactment'. Available at: http://arxiv.org/abs/2202.12972.
Perov I, Gao D, Chervoniy N, Liu K, Marangonda S, Umé C, Dpfks Mr, Facenheim CS, RP L, Jiang J, Zhang S, Wu P, Zhou B, Zhang W (2020) DeepFaceLab: Integrated, flexible and extensible face-swapping framework. http://arxiv.org/abs/2005.05535
Rankred (2022) 8 Best Deepfake Apps and Tools In 2022, Rankred. Available at: https://www.rankred.com/best-deepfake-apps-tools/ (Accessed: 11 December 2023).
Raza A, Munir K, Almutairi M (2022) 'A novel deep learning approach for deepfake image detection', Applied Sciences (Switzerland), 12(19)https://doi.org/10.3390/app12199820
Sabir E, Cheng J, Jaiswal A, AbdAlmageed W, Masi I, Natarajan P (2019) Recurrent convolutional strategies for face manipulation detection in videos. http://arxiv.org/abs/1905.00582
Shahzad HF et al (2022) 'A review of image processing techniques for deepfakes', Sensors. MDPIhttps://doi.org/10.3390/s22124556
Siegel, D. et al. (2021) 'Media forensics considerations on deepfake detection with hand-crafted features',. J ofImaging, 7(7). 10. 3390/jimaging7070108
Silva SH et al (2022) 'Deepfake forensics analysis: an explainable hierarchical ensemble of weakly supervised models', Forensic Science International: Synergy, 4https://doi.org/10.1016/j.fsisyn.2022.100217
van der Sloot B and Wagensveld Y. (2022) 'Deepfakes: regulatory challenges for the synthetic society', .Comput Law Secur Rev, 46https://doi.org/10.1016/j.clsr.2022.105716
Somers, M. (2020) Deepfakes, explained, MIT Management Sloan School. Available at: https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained (Accessed: 4 February 2023).
Uddin Mahmud B, Sharmin A (2020) Deep insights of deepfake technology : a review (Vol. 5, Issue 2). https://doi.org/10.48550/arXiv.2105.00192
Tolosana R, Romero-Tapiador S, Vera-Rodriguez R, Gonzalez-Sosa E, Fierrez J (2022) Deepfakes detection across generations: analysis of facial regions, fusion, and performance evaluation. Eng Appl Artif Intell 110. https://doi.org/10.1016/j.engappai.2022.104673
Wen D, Han H and Jain A K (2015) Face spoof detection with image distortion analysis, IEEE TRANSACTIONS ON INFORMATIONFORENSICS AND SECURITY. Available at: https://support.apple.com/kb/SP670
Wilpert C (2022) 7 best deepfake software apps of 2022 (50 Tools Reviewed), content mavericks. Available at: https://contentmavericks.com/best-deepfake-sof tware/ (Accessed: 24 December 2022).
Xu Z, Hong Z, Ding C, Zhu Z, Han J, Liu J, Ding E (2022). MobileFaceSwap: a lightweight framework for video face swapping. www.aaai.org
Yang T et al. (2022) 'Deepfake network architecture attribution', arxiv [Preprint]. Available at: https://arxiv.org/abs/2202.13843
Yousif, N. (2024). US FCC makes AI-generated robocalls illegal. https://www.bbc.co.uk/news/world-us-canada-68240887
Zhang Y, Zheng L, Thing VLL (2017) 2017 IEEE 2nd International Conference on Signal and Image Processing. ICSIP, Singapore. https://doi.org/10.1109/SIPROCESS.2017.8124497
Zichichi M et al (2022) Data governance through a multi-DLT architecture in view of the GDPR. Clust Comput 25(6):4515–4542. https://doi.org/10.1007/s10586-022-03691-3