Artificial Intelligence and Social Computing, Vol. 72, 2023, 216–224
https://doi.org/10.54941/ahfe1003290

Understanding Deepfakes: A Comprehensive Analysis of Creation, Generation, and Detection

Sami Alanazi and Seemal Asif

Centre for Robotics and Assembly, Cranfield University, Cranfield, MK43 0AL, ENGLAND

ABSTRACT

This paper provides a comprehensive analysis of deepfakes, focusing on their creation, generation, and detection. Deepfakes are realistic fabricated videos, images, or audios generated using artificial intelligence algorithms. While initially seen as a source of entertainment and commercial applications, the negative social consequences of deepfakes have become apparent. They are misused for creating adult content, blackmailing individuals, and spreading misinformation, leading to a decline in trust and potential societal implications. The paper also discusses the importance of legislation in regulating the use of deepfakes and explores techniques for their detection by using machine learning. Understanding deepfakes is essential to address their ethical and legal implications in today's digital landscape.

Keywords: Deepfakes, Artificial intelligence, Autoencoders, Deep neural networks, Detection algorithms, Generative adversarial networks, Fake content, Image manipulation

INTRODUCTION

Deep fake images or videos are content that is fake but looks original with the help of artificial intelligence algorithms. Such content can be technically proven wrong, whereas the naked eye cannot easily prove that. A mixture of "deep learning" and "fake" films are digitally altered hyper realistic videos which portray individuals who say and do things that have never happened genuinely. In particular, deep fake content is produced by aligning the faces of two different persons and then training auto-encoder to learn some features of one face, which as designated as "face A" and incorporate those into another face, referred to as Face B, and then generate a face that looks like B, but is not representative of their actual appearance. The reconstruction of faces based on selected features is used in black markets to create adult or related content for malpractices. Deepfakes depend on neural networks which scan a huge amount of data samples to learn to replicate a human's face, mannerisms, and voice, thus might cause serious consequences because it is very hard for people to distinguish them. Furthermore, it does not require experts to make realistic-looking sounding fake content. This is because non-experts can generate these types of Deepfakes artifacts by using readily available tools such as Face2Face and FaceSwap.

Unfortunately, Deepfakes are commonly used for negative purposes, including criminal activities such as impersonating the voice of a businessperson or employing them in political and pornographic contexts.

In light of these difficulties, it is crucial to investigate the use of detection algorithms and important methodologies to address the potential risks associated with deepfake technology. This review research paper aims to provide a thorough analysis of the creation and detection of deepfakes and contribute to a deeper understanding of this concerning technology.

DEEPFAKE GENERATION

Deepfakes are created using deep neural networks, specifically autoencoders. This process involves training a neural network to encode and decode images or videos as illustrated in Figure 1. The encoder takes the input image or video and compresses it into a latent code, which contains the essential features while removing unnecessary details. This latent code is then passed to the decoder, which recreates the original content from the code (Nguyen *et al.*, 2019).

To generate deepfake content, the autoencoder is trained on pairs of real and fake videos or images. The encoder learns to encode both real and fake content to produce similar latent codes. Meanwhile, the decoder reconstructs the original input from the encoded fake latent code, resulting in a realistic-looking deepfake content.

Deepfake content generation relies on various technologies, such as 3D ResNet and 3D ResNeXt algorithms. The proliferation of manipulated images and videos highlights the need for robust detection methods to distinguish between real and fake content. (Yang et al., 2022) propose Deepfake Network Architecture Attribution, which attributes fake images to their respective generator architectures. Their method performs well even with sophisticated models retrained on different datasets. Figure 2 demonstrates that architecture-level attribution is coarse-grained, while model-level attribution is relatively fine-grained. The evaluation compares two methods, learned features and AttNet, for extracting distinct features from GAN-generated images. While both methods are effective on the same set of GAN models and

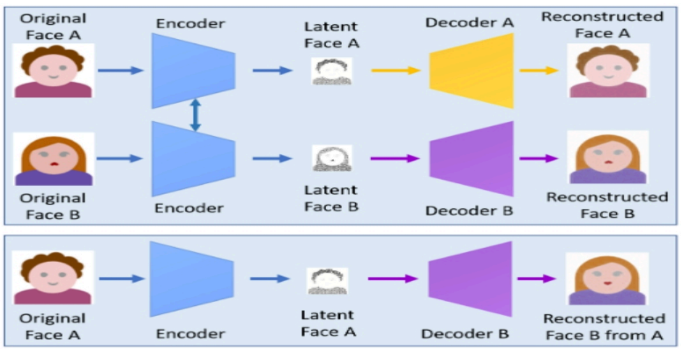

Figure 1: A deepfake model with two pairs of Encoder-Decoder (Nguyen et al., 2019).

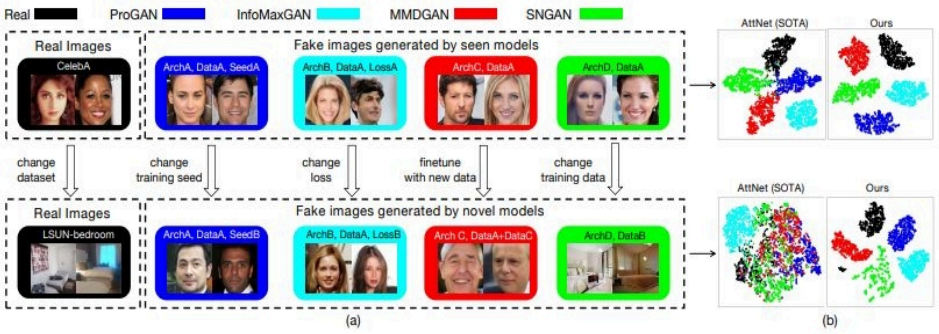

Figure 2: Deepfake generation (Yang et al., 2022).

real training images, AttNet fails to extract distinct features when tested on novel or differently trained images, unlike the proposed method (Yang et al., 2022). The figure also includes a t-SNE visual comparison of the learned features from both methods.

Software and Apps for Deepfake

The rapid evolution of deepfake creation technology, driven by the demand in black markets, necessitates the continuous improvement of detection technologies (Shahzad et al., 2022). Several tools are currently employed for the generation of deepfake content, which are outlined below.

One widely used tool, DeepSwap, for creating deepfake content in a recreational context is considered to be user-friendly and easily accessible online. Its free version is particularly popular among users, who can install the application on their mobile devices or utilize it on their laptops. The tool boasts two notable features. Firstly, it exhibits exceptional speed in generating output, enabling the creation of deepfake content in a remarkably short time (Wilpert, 2022). The processing efficiency of the tool ensures quick results. Additionally, the generated images bear a striking resemblance to genuine ones, posing a challenge for observers to discern between real and fake content at first glance (Rankred, 2022). Nevertheless, users have raised concerns regarding the difficulty of unsubscribing from the application, with the tool seemingly making it arduous for users to terminate their subscription. Consequently, only a limited number of users recommend the tool within their social circles.

DeepFace Lab is a platform widely used by students and researchers for creating manipulated images and videos on computers. While it may not be as user-friendly for the general public, researchers appreciate its flexibility in choosing the machine learning technology used (Wilpert, 2022). The interface is simple, but researchers with programming skills find it valuable. Moreover, the application is compatible with computers of varying processing power, making it accessible to a wider range of users (Rankred, 2022).

DeepFace Lab excels in generating high-quality output and offers an open-source platform for replacing the head and face of individuals in original images. Another notable feature is the ability to de-age faces in provided

images. Models and actresses can also benefit from the application, although the lack of a user-friendly interface may present challenges for them.

Deep Nostalgia is a popular deepfake app that generates high-resolution images and videos resembling genuine ones. It offers sharp edges, a photo enhancer for clear images, and appeals to users interested in fun content creation and sharing nostalgic animated representations.

Deep Art Effects is available for both computers and mobile devices, but mobile users are generally dissatisfied with the output. Compatibility issues with mobile phones, including the iPhone, are reported. The free version receives low ratings, while the paid version is considered better. Refund difficulties and inconvenient image selection contribute to its low popularity as a deepfake tool (Wilpert, 2022).

This web-based tool is only compatible with computers, not mobile devices. It has a steep learning curve and slow processing time. There are free and paid versions with trade-offs in quality and user-friendliness. Users should choose a tool that suits their needs (Wilpert, 2022; Rankred, 2022).

Deepfake Detection

Deepfakes pose increasing threats to privacy, security, and democracy. With the emergence of this danger, methods to detect deepfakes have been proposed. Early attempts relied on identifying manufactured characteristics stemming from glitches and discrepancies in falsely synthesized videos. Recent approaches have leveraged deep learning to extract significant and discriminatory features for detecting deepfakes (Chesney and Citron, 2019).

Deep detection is typically approached as a binary classification problem, distinguishing between authentic and manipulated videos. However, this process requires a large dataset of genuine and counterfeit videos to train classification models (de Lima et al., 2020). Despite the increasing availability of fake videos, there is a lack of standardized benchmarks for evaluating different identification techniques. To address this challenge, Korshunov and Marcel (2018) developed a notable deep dataset comprising 620 video models generated using the Faceswap-GAN open-source code. The dataset utilized publicly available movies from the VidTIMIT database to create deepfakes with realistic facial expressions, mouth movements, and eye blinks. These videos were then used to evaluate various detection techniques.

The test findings reveal that popular facial recognition systems based on VGG and Facenet are not successful in detecting deepfakes. Additionally, methods such as lip-syncing and picture quality measurements using Support Vector Machines (SVMs) exhibit a significantly high error rate when applied to identify deepfake videos in this newly generated dataset. These results raise concerns about the urgent need to develop more robust approaches for detecting deepfakes (Wen, Han and Jain, 2015). The following sections will define different types of deepfake detection methods.

Fake Image Detection

Face-swapping has numerous appealing applications in video composition, portraiture, and identity protection, allowing the substitution of faces in

images with those from a collection of photographs. However, cybercriminals have also been employing these techniques to infiltrate systems and gain illegal access for identity theft or unauthorized authentication (Korshunova et al., 2017). The use of deep learning methods like Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) has significantly increased the difficulty of detecting swapped facial images for forensic models, as they can preserve the facial position, expressions, and lighting. (Zhang, Zheng and Thing, 2017) utilized a word bag technique to extract a collection of compact features, which were then fed into various classifiers such as Support Vector Machines (SVMs) and Multi-Layer Perceptron (MLPs) to differentiate between real and swapped face images. Among different types of deepfake images, GAN models that have been synthesized are particularly challenging to recognize due to their high quality, realism and the GAN's ability to model complex data distributions and generate outputs that match the input distribution.

As a hypothesis testing problem, (Agarwal and Varshney, 2019) considered GAN deep detection to be a statistical framework, based on the information-theoretical authentication research. They determined the minimum distance, termed the oracle error, between the distribution of valid pictures and the images generated by a specific GAN. The analytical findings demonstrate that as the precision of the GAN decreases, this distance increases, making it easier to detect profound defects in deepfakes. In the case of high-resolution image inputs, GANs are necessary to generate counterfeit pictures that are challenging to identify (Nguyen et al., 2019).

Fake Video Detection

Due to significant deterioration of frame data after video compression, most methods of image identification are not suitable for films. Moreover, videos feature temporal characteristics that vary from the frameset to a methodology that only fakes pictures can be detected (Afchar et al., 2018). Deepfakes techniques of video can be categorized into two groups:

Video Frames Temporal Features: (Sabir et al., 2019) exploited video stream Spatio-temporal properties to detect depth defects on grounds that time coherence is not effectively imposed on the deepfakes synthesis process. Video modification is carried out on an interface-by-frame basis to further exhibit low-level anomalies caused by facial changes as temporal objects with contradictions between frames. The process for detecting face manipulation involves two steps. The first step involves detecting, cropping, and aligning faces on a sequence of frames. The second step uses a combination of a convolutional neural network (CNN) and a recurrent neural network (RNN) to differentiate between manipulated and authentic face images as Figure 3 explains (Nguyen et al., 2019). The suggested process is evaluated on the 1000-video FaceForensics++ data set, showing encouraging results.

Video Frame Visual Artifacts is a method that analyzes individual frames of a video to identify visual characteristics for distinguishing between real and deepfake videos. Meso-4, introduced by (Afchar *et al*., 2018), is a deep learning technique that employs a complex architecture with convolutional

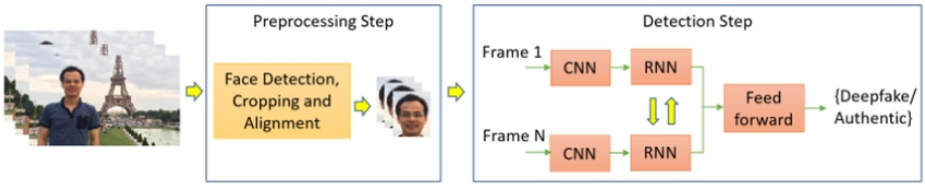

Figure 3: Face manipulation detection two-step process (Nguyen et al., 2019).

and pooling layers to detect deepfake elements. MesoInception-4 is an enhanced version of Meso-4 that incorporates the inception module for increased model optimization. While Meso-4 has advantages such as conducting binary classification and identifying deepfake and real images, it relies on a shallow CNN architecture, which may limit its ability to detect sophisticated manipulations. Neural networks are effective in deepfake detection, with a focus on face warping artifacts and physiological/biological features. (Raza, Munir and Almutairi, 2022) propose a deepfake detection model that employs neural network techniques trained on a dataset of fake and real human faces, achieving high accuracy in identifying deepfake elements.

The use of deepfake datasets, obtained from sources like Kaggle, enables the training and testing of neural network techniques for deepfake detection. Transfer learning-based models are employed, utilizing pre-trained models to predict real and fake images by analyzing facial features. Algorithms analyze dimensions, size, and shapes of features to identify patterns and classify images or videos as fake if inconsistencies are detected. The Xception Technique, a transfer learning-based neural network method, utilizes deep separable convolution layers to detect alterations in images and videos. Different deepfake detection techniques may outperform others based on factors like dataset size or algorithm sophistication. Techniques such as pro-3D CNN and physiological measurements like heart rate using long-distance photoplethysmography (rPPG) show promise but require further development. Meta-learning techniques are also being explored for deepfake detection. The current complexity and time-consuming nature of forensic processes highlight the need for more efficient applications to certify and verify the authenticity of videos and images. Deep learning methods offer significant potential in distinguishing between fake and authentic content, but further advancements are necessary to address the challenges posed by deepfake technology.

Manipulating Images/Videos With Human Expressions in Deepfakes

The manipulation of static images is relatively simpler compared to that of moving images. However, altering videos involving human expressions poses a significant challenge in deepfake content manipulation. Each individual possesses unique styles of expressions that, when combined with their facial features, yield distinct visual outcomes. According to (Groh et al., 2021), deepfake videos are typically generated from open-source datasets, wherein human faces appear akin to puppets devoid of any discernible expressions. To

address this limitation, advanced deepfake technologies have emerged, focusing on manipulating movements encompassing facial and body motions and expressions. Artificial intelligence is employed to model human behaviors like walking, talking, smiling, crying, and frowning, which are subsequently utilized to overwrite the original identity. Notably, videos of shorter duration, featuring fewer expressions, are relatively easier to manipulate than those characterized by complex expressions, numerous expressions, and longer temporal extents.

Sophisticated algorithms employ psychology, probability, kinematics, inverse kinematics, and physics to detect deepfake content by analyzing temporal aspects of videos. Face-centric neural network algorithms (CNN) are considered highly accurate for deepfake detection, focusing on facial location rather than emotion-congruent speech and expressions (Groh et al., 2021).

Detecting deepfake manipulation involves analyzing specific facial regions rather than the entire image. Algorithms use fusion techniques to detect tampering by comparing these regions with a large training set that captures facial characteristics across different demographics. Random labels like facial expression, hair, and eyes are used to assess changes, as tiny distortions in facial regions, though imperceptible to humans, can significantly impact the output image. Algorithms focus on monitoring these selected regions for accurate detection (Tolosana et al., 2022; Guarnera et al., 2022).

Detecting deepfake content involves considering the background and scene elements, not just the person depicted. Algorithms are trained to identify changes in scenes by starting with simple backgrounds and gradually increasing complexity. Rotation of scene elements and input from domain experts help identify critical features associated with specific scenarios. By detecting alterations or changes in these features, algorithms can label deepfake images based on the detected tampering (Choras et al., 2020; Siegel et al., 2021).

Data scientists and AI specialists are exploring methods to detect fake images and videos by analyzing both prominent features like accents and subtle details such as lighting. Training sets focus on poses, postures, lighting conditions, and backgrounds to determine authenticity. The natural physics of lighting holds promise in identifying deepfakes, but AI tools have yet to fully harness this domain. Ongoing research investigates the physics of lighting to enhance deepfake forensics (Somers, 2020).

Generative Adversarial Networks (GANs) are raising concerns about privacy and trust among internet users due to the creation of hyper-realistic deepfakes. GANs enhance manipulated images through adversarial and perceptual losses, resulting in visually convincing forgeries. Frame-to-frame face detection and facial reenactment techniques further improve the realism of GAN-processed videos. Face morphing and face swap are common deepfake techniques, with face morphing involving blending features from multiple individuals. Detecting morphed facial images is crucial for reliable recognition systems, and techniques such as morphing attack detection (MAD) can be used. GANs play a significant role in data forgery and image manipulation, generating high-resolution fake images that are challenging to distinguish from real ones. Techniques like deep convolutional

generative adversarial networks (DCGAN) aid in training GANs to produce more deceptive images accurately.

Phoneme-viseme mismatches, where the sound of a letter or alphabet does not align with the shape of the mouth, are used to detect deepfake videos (Agarwal *et al.*, 2020). These small yet significant inconsistencies can help identify manipulations. Language specialists are consulted to detect deepfakes in different languages. Forensic techniques involving human involvement are utilized, where deep learning algorithms aid in decision-making. Attention-based explainable deepfake detection algorithms allow experts to focus on specific regions of images and videos. Human instincts and cultural context play a role in detecting deepfakes. Manual selection of regions by forensic experts can be further processed using software for accurate detection.

DISCUSSION AND CONCLUSION

The rapid progression of deepfake technology has raised concerns about its potential for deception and misuse. Laws are being enacted to protect internet users and create a secure cyber space. Detecting deepfake content has become a challenge, but researchers have identified characteristics such as unnatural eye blinking patterns that can be used for detection. Deepfake systems initially lacked realistic blinking, but newer techniques have incorporated this feature. Detecting deepfakes can be complex and requires training machines to recognize variations in blinking patterns for different individuals and situations. AI and other technologies are being used to improve the detection and prevention of deepfake content. While visual differences between real and fake content can be subtle, machine learning algorithms can identify discrepancies in eye blinking and facial expressions. These advancements highlight the importance of relying on technology rather than solely relying on human observation to detect deepfake content.

In conclusion, deepfake technology presents both benefits and risks. Policy making is necessary to address the dangers of deepfake content, including state-level laws, social media platform policies, and national laws to punish those who create and share deepfakes for malicious purposes. Public awareness campaigns should educate society about the ethical boundaries of deepfake creation and sharing. Collaboration between government, technology companies, and society is essential in developing techniques to detect and prevent deepfakes. Cyber law enforcement needs to evolve to ensure the security of all users in the online space. Continued innovation and regulation are crucial in addressing the challenges posed by deepfakes.

REFERENCES

Afchar, D. *et al.* (2018) 'MesoNet: a Compact Facial Video ForgeryDetection Network'. Available at: https://doi.org/10.1109/WIFS.2018.8630761.

Agarwal, S. *et al.* (2020) *Detecting Deep-Fake Videos fromPhoneme-Viseme Mismatches*. Available at: www.instagram.com/bill_posters_uk.

Agarwal, S. and Varshney, L. R. (2019) 'Limits of Deepfake Detection: ARobust Estimation Viewpoint'. Available at: https://arxiv.org/abs/1905.03493.

Chesney, B. and Citron, D. (2019) 'Deep fakes: A looming challenge for-privacy, democracy, and national security', *California LawReview*, 107(6), pp. 1753–1820. Available at: https://doi.org/10.15779/Z38RV0D15J.

Choras, M. *et al.* (2020) 'Advanced Machine Learning Techniquesfor Fake News (Online Disinformation) Detection: A Systematic MappingStudy', *Applied Soft Computing* [Preprint]. Available at: https://arxiv.org/abs/2101.01142.

Groh, M. *et al.* (2021) 'Deepfake Detection by Human Crowds, Machines, and Machine-informed Crowds', *arxiv* [Preprint]. Available at: https://doi.org/10.1073/pnas.2110013119.

Guarnera, L. *et al.* (2022) 'The Face Deepfake DetectionChallenge', *Journal of Imaging*, 8(10). Available at: https://doi.org/10.3390/jimaging8100263.

Korshunov, P. and Marcel, S. (2018) 'DeepFakes: a New Threat to FaceRecognition? Assessment and Detection'. Available at: https://arxiv.org/abs/1812.08685.

Korshunova, I. *et al.* (2017) 'Fast Face-swap Using ConvolutionalNeural Networks'. Available at: https://arxiv.org/abs/1611.09577.

de Lima, O. *et al.* (2020) 'Deepfake Detection usingSpatiotemporal Convolutional Networks'. Available at: https://arxiv.org/abs/2006.14749.

Nguyen, Thanh Thi *et al.* (2019) 'Deep Learning for DeepfakesCreation and Detection: A Survey'. Available at: https://doi.org/10.1016/j.cviu.2022.103525.

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. and Almutairi, M. (2022) 'A Novel Deep LearningApproach for Deepfake Image Detection', *Applied Sciences (Switzerland)*, 12(19). Available at: https://doi.org/10.3390/app12199820.

Shahzad, H. F. *et al.* (2022) 'A Review of Image ProcessingTechniques for Deepfakes', *Sensors*. MDPI. Available at: https://doi.org/10.3390/s22124556.

Siegel, D. *et al.* (2021) 'Media forensics considerations ondeepfake detection with hand-crafted features', *Journal ofImaging*, 7(7). Available at: https://doi.org/10.3390/jimaging7070108.

Somers, M. (2020) *Deepfakes, explained*, *MIT ManagementSloan School*. Available at: https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained (Accessed: 4 February 2023).

Wen, D., Han, H. and Jain, A. K. (2015) *Face Spoof Detection withImage 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 ToolsReviewed)*, *contentmavericks*. Available at: https://contentmavericks.com/best-deepfake-software/ (Accessed: 24 December 2022).

Yang, T. *et al.* (2022) 'Deepfake Network ArchitectureAttribution', *arxiv* [Preprint]. Available at: https://arxiv.org/abs/2202.13843.

Zhang, Y., Zheng, L. and Thing, V. L. L. (2017) *2017 IEEE 2nd International Conference on Signal and Image Processing, ICSIP : August 4-6, 2017, Singapore*.

2023-07-24

Understanding deepfakes: a comprehensive analysis of creation, generation, and detection

Alanazi, Sami

AHFE International

Alanazi S, Asif S. (2023) Understanding deepfakes: a comprehensive analysis of creation, generation, and detection. In: 14th International Conference on Applied Human Factors and Ergonomics (AHFE 2023) and the Affiliated Conferences, 20-24 July 2022, San Francisco, USA
https://doi.org/10.54941/ahfe1003290