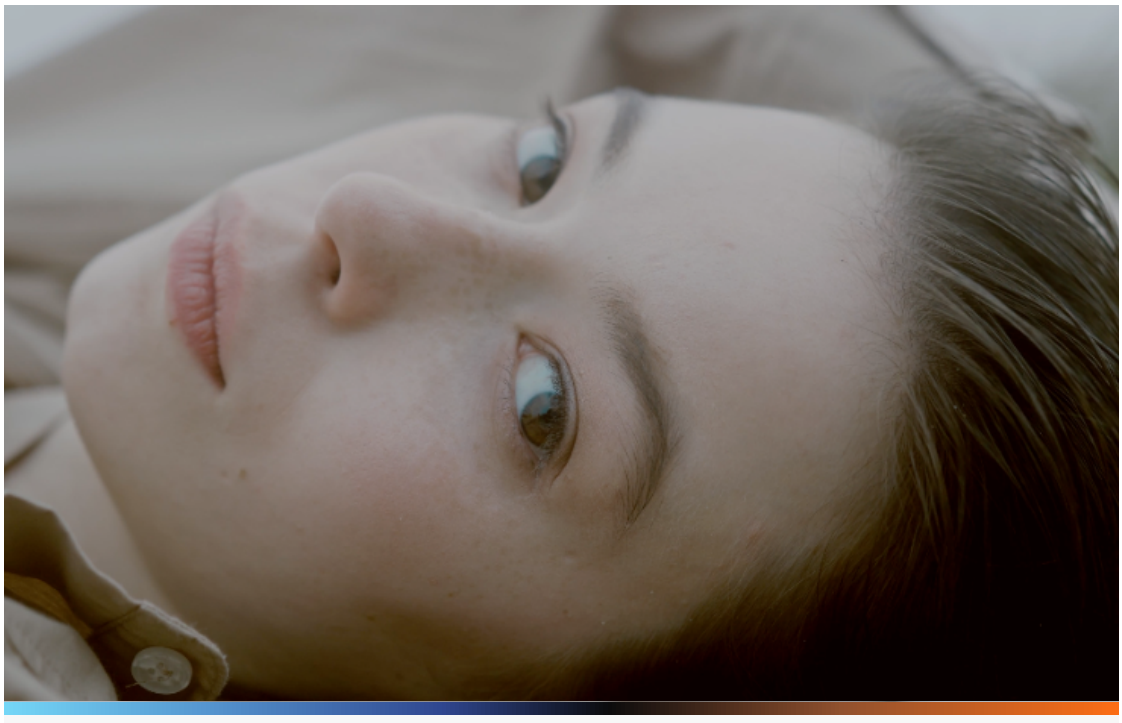
PARAVISION

A PRACTICAL GUIDE TO **DEEPFAKE DETECTION**

Definitions, Challenges, and Solutions



paravision.ai

Trusted Identity AI

October 2024

Introduction

In this era of unprecedented technological progress, the increasing use of deepfaked imagery in traditional and social media is ushering in new challenges, particularly in identity, privacy, and trust. As the creation and consumption of hyper-realistic deepfake-powered imagery is increasingly pervasive, the spread of digitally manipulated imagery is undermining the public's understanding of what is real and what is not. This has significant implications for democracy, national security, and human rights. Meanwhile, in the context of digital identity, deepfakes are emerging as a threat in both selfie and identity document images, especially in remote enrollment scenarios.

While deepfakes are thoroughly discussed by companies, academia, the press, and the general public, the topic of deepfakes is complex, multifaceted, and frequently misunderstood. In this guide, we will introduce the landscape of facial deepfakes and deepfake detection, covering the technical complexities of what deepfakes are, the threats they pose, the role of AI in detection, and the significance of accurate identification in preserving the integrity of digital content.

This guide focuses specifically on the detection and prevention of **deepfakes involving faces in imagery and video**, and does not cover deepfakes related to audio manipulation or general image tampering.

© Paravision 2024

Table of Contents

Understanding Deepfakes **4**

Types of Digital Face Manipulations 7

Examples of Digital Face Manipulations 8

Challenges Posed by Deefakes **9**

Detecting Deepfakes **10**

Human Detection of Deepfakes 10

Detecting Deepfakes with AI 10

Human-in-the-Loop, Powered by AI 11

Critical Metrics for Assessing Deepfake Detection Technology 12

Best Practices for Deepfake Detection **13**

Paravision Deepfake Detection **15**

Conclusion **16**

© Paravision 2024

Understanding Deepfakes

Deepfake technologies leverage AI to create hyper-realistic synthetic media, including manipulated imagery, video, and audio, in which an identity is presented in an inauthentic context. These technologies can be used to convincingly manipulate media and change its content while making it seem authentic to human perception. Compared to more traditional image or video manipulation, deepfake technologies are increasingly realistic and challenging for humans to detect.

Misunderstandings about deepfakes are common because of the rapidly evolving technology space and the natural overlap with other digital face manipulations, such as synthetic faces and face morphs (which also may be used to undermine identity in various contexts). For example, it could reasonably be argued that the Oxford English Dictionary's 2023 definition of deepfake is already outdated. Per the OED, the definition for deepfake is:

> *Any of various media, esp. a video that has been digitally manipulated to replace one person's likeness convincingly with that of another, often used maliciously to show someone doing something that he or she did not do.*[1]

However, this doesn't cover the context of a fully synthetic image (e.g., created with a diffusion technology like Stable Diffusion) that convincingly and maliciously shows an individual doing something they did not do. A synthetic image is not a digitally manipulated image that replaces one person with another, but the use and outcome can be similar enough to suggest it fits the description of a deepfake. And if not exactly a deepfake, it is still relevant to the problem statement surrounding digital face manipulations. Here, we will refer to such an image as a deepfake since the goal and result are deeply related to more "traditional" deepfake technologies. Importantly, in this guide we don't cover deepfakes related to audio manipulation or general image tampering, but focus specifically on the detection and prevention of deepfakes involving faces in imagery and video.



1. https://www.oed.com/dictionary/deepfake_n

To understand face-related deepfakes, we need to articulate the variety of ways they can be manifested in digital media. Some common digital face manipulations include:

Face Swaps (a.k.a. Identity Swaps)	Face Swaps involve seamlessly replacing a person's face in a video, mimicking their expressions and movements to create a convincing impersonation. Consider the scenario of a video call in which a CEO's face is swapped onto the head of a similar-looking actor in order to make false claims about the company.
Expression Swaps	Expression swaps, sometimes called "puppet master" deepfakes, involve using an image or video of one person and manipulating the movements and facial expressions. Consider the scenario of an existing video of the president of one country manipulated to seem as if they are declaring war on another country.
Adaptation and Editing	Deepfake technologies can be used to digitally manipulate or add visual attributes to an image or video. Consider adding a scar, tattoo, or other feature to a photo of a celebrity and giving it to the press.
Text-to-Image, Text-to-Video	With the latest diffusion-based imagery technologies, it is possible to enter a line (or block) of text and create an image or even a video of known individuals in a fully synthetic situation. Technologies like DreamBooth (or more generically LoRA) make it easy to insert any individual into such a scheme without much time, cost, or effort.
AI Avatars	Deepfakes are also used to create lifelike talking head videos where the subject appears to speak naturally, with synchronized lip movements and facial expressions. Unlike face swaps, AI Avatars don't require an actor's face to be overlaid; instead, they animate the original subject's face, head, or body directly, using minimal video footage as a reference.

Even though not technically deepfakes, the following digital face manipulation technologies can also be used to subvert identity:

Face Morphing	Facial morphs involve combining two facial images into a new face using artificial intelligence. The result is a face that could pass as either one of the original individuals in identity verification scenarios. This is particularly concerning in the case of identity documents (like passports) based on insecure analog (printed) photos.
Synthetic Faces	Synthetic faces are entirely artificial faces generated by deep learning models, often with no real-world counterpart. They can be the foundation of identity fraud at scale and have been seen by the thousands on social networks such as LinkedIn. Synthetic faces were first made popular with "This Person Does Not Exist" and StyleGAN2 but are now inherently part of text-to-image, diffusion-based image generators.
Image Perturbation	Image perturbation is a type of adversarial attack that involve manipulating the image to deceive a machine learning model in a way that is not visually obvious. This approach is more conceptual/academic at this point but is cause for concern as AI technologies become more advanced.

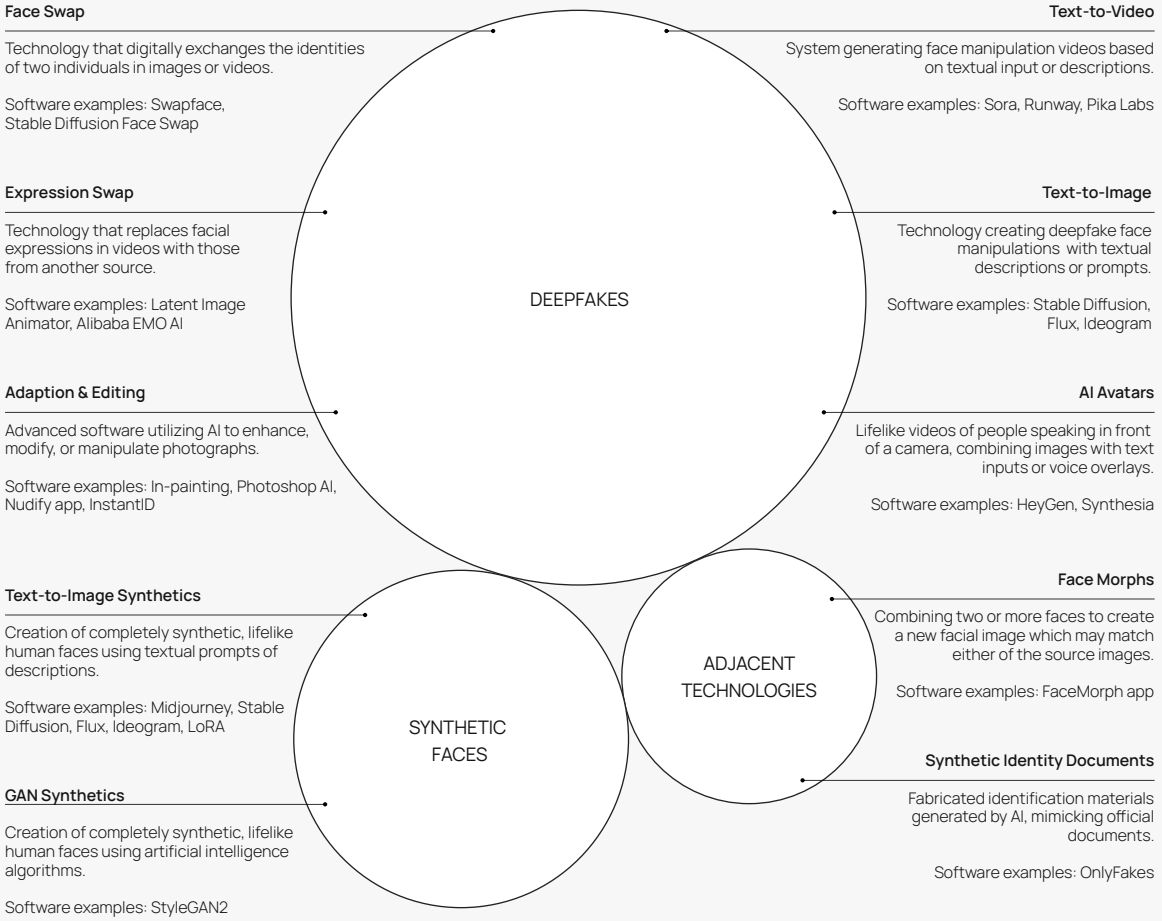
Deepfakes can also be classified by the type of AI technology used for their creation. Facial Deepfakes are often made using one of the two prominent technologies: Generative Adversarial Networks (GANs) and diffusion models.[2] While both aim to generate high-quality imagery, they differ significantly in their methodologies and underlying architectures.

GANs (Generative Adversarial Networks)	GANs operate on the principle of an adversarial game between a generator and a discriminator. This involves the generator producing samples to deceive the discriminator, which aims to distinguish between real and fake data. Through this iterative process, GANs excel at generating highly realistic and diverse samples, showcasing details across various mediums like images, audio, and text. However, GANs face challenges such as mode collapse, where they may produce repetitive or limited samples, necessitating a delicate balance in training to ensure reliable results. GANs are commonly used for face swap (i.e., identity swap) and expression swap deepfakes.
Diffusion Models	In contrast, diffusion models take a different path to data generation. Operating through an iterative diffusion process, these models transform random noise into desired data by gradually refining it through successive steps. While diffusion models offer fine-grained control over the generation process and stability during training, they require longer training times and are computationally intensive. Nevertheless, they excel in tasks like image inpainting, denoising, and data synthesis, enabling the creation realistic samples with complex dependencies and patterns. Whereas GANs are most frequently seen in traditional deepfakes, diffusion models have rapidly become a major threat due to their realistic outputs and ready access online.
Neural Talking Heads	Neural Talking Heads technology uses a combination of Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Generative Adversarial Networks (GANs), Transformers, and other deep learning models to animate human faces from reference images or videos. By training on large datasets of face videos, these systems learn spatial and temporal patterns of facial movements, allowing them to animate faces convincingly based on new audio or target expressions. This technology is widely used in synthetic media, enabling applications like AI avatars, digital puppeteering, and deepfakes where precise facial animation is critical. Neural Talking Heads are functionally a next-generation Expression Swap, but are much more visually compelling than GANs alone due to the layers of technology used.

2. https://www.sabrepc.com/blog/Deep-Learning-and-AI/gans-vs-diffusion-models

REFERENCE:

THE MOST COMMON TYPES OF FACE DEEPFAKES AND RELATED DIGITAL FACE MANIPULATIONS



Examples of Common Digital Face Manipulations:

DEEPFAKES

SYNTHETIC FACES

ADJACENT TECHNOLOGIES

Type	Real life example of manipulation — Visual example
Face Swap	In 2024, a company in Hong Kong fell victim to an elaborate scam, where the company's CFO took part in a video meeting and was convinced to pay $25M to fraudsters using identity swap to portray as other company leadership.[3] Another famous example of identity swap are the Tom Cruise deepfakes that surfaced in 2021.[4]
Expression Swap	In 2022, a manipulated video circulated on social media purportedly showing Ukrainian President Volodymyr Zelenskiy publicly capitulating to Russian demands.[5] The video used expression swap to "puppet-master" a real image of President Zelenskiy with a video of someone else speaking the words heard in the video.
Adaption & Editing	AI-based filters and editing tools have made it difficult to detect real from fake on any social media platforms. In 2024, AI Nudifying apps were used in multiple middle and high schools by students to create fake nude photos of their classmates, leading to police investigations.[6]
Text-to-Image, Text-to-Video	In 2023, an image created on Midjourney showing Donald Trump being arrested by the NYPD was circulated on social media, causing concern on spread of misinformation.[7] In 2024, text-to-image technology was used to create explicit deepfakes of Taylor Swift, causing the social media platform X to block searches with her name.[8]
AI Avatars	In 2024, Reid Hoffman published a YouTube video where he interviewed his new digital avatar, Reid AI, using this technology.[9] Companies like HeyGen and Synthesia leverage this approach to create realistic avatars that can communicate in any language, making it a powerful tool for use on a global scale.
Synthetic Faces	Fully synthetic faces created with technologies such as StyleGAN2 can have been seen by the thousands on social networks such as LinkedIn. Synthetic faces and profiles can be used to share disinformation, promote political messages, and conduct mass outbound sales activities by companies.[10]
Face Morphs	Facial morphs have uses in entertainment use cases, such as morphing two celebrities' faces into one. Morphs have been used in identity documents, posing risks in facial authentication and government identity applications, and causing concern that morphs could pass for either of the source identity in digital biometric checks.[11]
Synthetic ID Documents	Identity providers have reported significant rises in AI technologies being used to create synthetic or forged ID documents. Deepfake imagery in document authentication posed challenges for identity theft and fraud, especially in remote and self authentication settings.[12]

3. https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html
4. https://www.tiktok.com/@deeptomcruise?lang=en
5. https://www.reuters.com/world/europe/deepfake-footage-purports-show-ukrainian-president-capitulating-2022-03-16/
6. https://www.nytimes.com/2024/04/08/technology/deepfake-ai-nudes-westfield-high-school.html
7. https://www.bbc.com/news/world-us-canada-65069316
8. https://www.cbsnews.com/news/taylor-swift-deepfakes-x-search-block-twitter/
9. https://www.youtube.com/watch?v=rgD2gmwCS10
10. https://www.npr.org/2022/03/27/1088140809/fake-linkedin-profiles
11. https://www.businessinsider.com/artist-morphs-celebrities-faces-together-to-create-perfect-hybrids-2020-3
12. https://www.404media.co/inside-the-underground-site-where-ai-neural-networks-churns-out-fake-ids-onlyfake/

© Paravision 2024

Challenges Posed by Deepfakes



Challenges to Democracy and National Security

Deepfakes wield the power to undermine democratic processes and destabilize national security through the dissemination of fabricated content. Instances of deepfakes manipulating political narratives and influencing elections have raised concerns globally. False claims propagated through deepfake-generated content can spread rapidly on social media platforms, challenging the integrity of democratic discourse.

Challenges to Businesses

Deepfakes threaten businesses, including financial institutions, security firms, and identity verification services. Malicious actors can exploit deepfakes for fraudulent activities, leading to economic losses and reputational damage. There have been reports of deepfakes being used to counterfeit job interviews with talent acquisition professionals on video conferencing platforms like Zoom and Teams, underscoring the potential for deepfakes to deceive and manipulate business operations.

Challenges to Identity

Deepfakes represent an escalating threat in digital identity verification, and their increasing sophistication makes them potent tools for identity fraud. Fraudsters can use deepfakes to bypass biometric security measures by attacking selfies, ID documents, or video communication channels within Identity Verification applications, pushing for advancements in detection technologies and new security frameworks to combat this evolving threat. Identity companies have reported an increase in deepfakes used in document and identity fraud, causing challenges for financial and identity companies. Fraudulent actors can use deepfake technologies, such as face swaps or synthetic faces, to present themselves as someone else or create manipulated or synthetic identity documents to access user accounts fraudulently.

Challenges to Human Rights and Personal Privacy

Deepfakes raise profound concerns regarding human rights violations and infringements on personal privacy, mainly through the unauthorized use of individuals' likenesses in fabricated content. The proliferation of deepfake-generated pornographic material disproportionately affects women and can have devastating consequences for victims. Studies indicate that 96% of deepfake videos on the internet are pornographic[13], highlighting the prevalence of non-consensual and malicious use of deepfake technology.

As deepfake technology evolves, these challenges underscore the urgent need for robust measures to mitigate the risks associated with synthetic media manipulation. Collaboration between governments, technology companies, and civil society is essential to develop effective strategies and safeguards to address the multifaceted threats posed by deepfakes.

13. https://www.nbcnews.com/tech/internet/deepfake-porn-ai-mr-deep-fake-economy-google-visa-mastercard-download-rcna75071

Detecting Deepfakes: Approaches and Metrics

Human Detection of Deepfakes

Detecting deepfakes poses a significant challenge due to their increasingly sophisticated nature. However, researchers and technology companies have developed various techniques to identify and mitigate the spread of synthetic media. Some of the prominent detection techniques include:

Forensic Analysis	Forensic analysis involves examining metadata, digital footprint, or other contextual information or insights to identify anomalies that may indicate manipulation. This technique can help uncover artifacts left behind during the deepfake generation process.
Visual Analysis	While detecting deepfakes visually with the human eye is becoming increasingly difficult, slight discrepancies between the synthesized content and genuine footage can be detected. Irregularities in facial expressions, eye movements, or body gestures can signal the presence of a deepfake or synthetic image. Deepfake imagery often struggles with human hand gestures, dental alignment, and eyewear, which can help detect deepfakes visually.
Multimodal Analysis	Multimodal analysis combines multiple detection techniques to enhance the accuracy of deepfake detection. By analyzing multiple media content modalities, identifying inconsistencies that may not be apparent in individual modalities alone becomes more accurate and trustworthy.

Detecting Deepfakes with AI

The technical prowess of AI in deepfake detection lies in its ability to analyze patterns and inconsistencies within media content. AI technologies play a pivotal role in scrutinizing facial features, expressions, and other image cues to identify anomalies indicative of deepfake manipulation, a task increasingly challenging for human detection with the advancement of technology.

AI-powered deepfake detection systems utilize sophisticated machine learning models to assess various aspects of media content, including facial feature analysis, voice analysis, pattern recognition, and behavioral analysis to detect the likelihood of an image or video being digitally manipulated. Trained on extensive datasets comprising authentic and manipulated images, these models leverage deep neural networks to extract features and patterns indicative of deepfakes.

Human-in-the-Loop, Powered by AI

AI-driven tools are indispensable in the arms race against increasingly realistic deepfakes, as they're uniquely capable of using advanced algorithms to analyze visual data at scale, identify inconsistencies, and flag potential fakes. These systems are trained on vast datasets and can spot subtle discrepancies often invisible to the human eye. For instance, they can detect irregular blinking patterns, unusual facial expressions, or inconsistencies in lighting that might indicate a manipulated video or image. This capability is crucial given the volume and sophistication of digital content.

However, while AI excels at processing and analyzing data quickly, it lacks the nuanced understanding necessary to grasp the full context behind every piece of content. This is where the human element becomes critical. Humans can consider cultural, social, and situational contexts that AI might overlook or misinterpret. Therefore, integrating human judgment into the decision-making process—human-in-the-loop—ensures a more balanced approach to action on AI recommendations. For example, before taking any significant actions based on AI flags, such as legal measures or public denouncements, human oversight can assess the broader implications and ethical considerations, ensuring that decisions are accurate and just. This synergy between human insight and AI capabilities is essential to navigate the challenges posed by deepfake technologies effectively, ensuring that AI aids but does not autonomously dictate actions that seriously affect individuals' lives or crucial business processes.



Critical Metrics for Assessing Deepfake Detection Technology

The accuracy of deepfake detection technologies is measured through various metrics, each addressing specific aspects of its performance:

Attack Presentation Classification Error Rate (APCER)	APCER is a critical metric used for liveness detection and presentation attack detection systems. This rate measures the likelihood that a biometric security system will incorrectly classify a spoofing attack (e.g., a fake photograph, a video, or a mask) as a legitimate biometric input from the genuine user. APCER is a crucial indicator of a biometric system's vulnerability to false acceptances of fraudulent attempts, quantifying the system's susceptibility to specific types of presentation attacks.
Bona Fide Presentation Classification Error Rate (BPCER)	BPCER is another important metric used to evaluate the performance of liveness and presentation attack detection systems. It quantifies the frequency at which a biometric system erroneously rejects an attempt as a spoofing attack when the biometric input is from a legitimate user. BPCER measures the system's tendency to identify genuine presentations as attacks mistakenly. A lower BPCER indicates a lesser likelihood of inconveniencing genuine users but might increase the risk of failing to detect actual attacks if tuned improperly.
APCER@BPCER=x	This metric assesses the APCER at a specific BPCER threshold, offering insights into model performance regardless of internal threshold parameters. It helps evaluate the effectiveness of deepfake detection systems across different thresholds.
Equal Error Rate (EER)	EER determines the threshold where APCER equals BPCER, providing a balanced performance indicator. Achieving a low EER indicates optimal performance in balancing false acceptance and rejection rates.

By leveraging AI technologies and robust performance metrics, deepfake detection systems can enhance their accuracy and reliability in identifying and mitigating the spread of synthetic media manipulation.

© Paravision 2024

Best Practices for Deepfake Detection

Detecting deepfakes requires a multifaceted approach that combines advanced technologies and vigilant monitoring strategies. To effectively combat the spread of synthetic media manipulation, here are some best practices for deepfake detection:



Take a Whole-of-Enterprise Approach

Deepfakes can undermine various aspects of business operations, from hiring processes to customer interactions and external communications. An integrated strategy across all departments ensures comprehensive monitoring and response mechanisms are in place. This approach includes:

- Assessing how deepfakes could impact hiring processes by presenting false identities or qualifications
- Evaluating the risks to live video and audio communications which could be manipulated to damage stakeholder trust or extract confidential information
- Implementing robust verification processes to safeguard customer identities and prevent fraud
- Ensuring that external marketing communications are protected from deepfake exploitation that could harm the brand's reputation



Use the Right Technology for Intended Use Case

As discussed throughout this paper, deepfakes are not “one thing.” It is critical to understand the specific use cases that are threatened (i.e. video communications vs. identity verification) and the specific threat vectors that are likely for these use cases (i.e. face swaps vs. diffusion images vs. GAN synthetics). Because the space is so broad and poorly defined, technology providers may legitimately provide solutions for one type of deepfake that is not the likely threat vector for the use case of concern, so close collaboration with providers is critical.



Consider Automated and Human-in-the-loop Analysis

Automated deepfake analysis can be incredibly powerful, finding signals and patterns in data that humans would not identify alone. At the same time, there are limits to deepfake detection technologies and the legitimate confusion about results even when correct. Proper combination of automated analysis and human review, context building, analysis, and communications helps to build the highest security and highest confidence system.



Implement Continuous, Live Monitoring

Deploy live monitoring systems to continuously monitor online platforms and social media networks to quickly detect and remove deepfake content. By leveraging automated flagging mechanisms and alert systems, you can mitigate the spread of malicious deepfake content before it reaches a wider audience.



Leverage Ethically Trained AI Models

Utilize AI models trained on ethically sourced and diverse datasets to improve the ethics, responsibility, and performance of your deepfake detection technology stack. By prioritizing using proprietary data that adheres to ethical guidelines and standards, you can enhance the detection algorithms' fairness, inclusivity, and accuracy. Ethically trained AI models mitigate the risk of bias and discrimination and reduce the organizational legal risk by staying ahead of AI regulations that will call for auditable, balanced, fair models. Additionally, incorporating diverse datasets ensures that detection systems are robust and effective across different demographics, enhancing their reliability and impact.



Collaborate with Industry Partners

Foster collaboration with industry partners, technology companies, and research institutions to share insights, resources, and best practices for deepfake detection. By leveraging collective expertise and resources, you can strengthen the effectiveness of detection efforts and stay ahead of emerging threats.



Educate and Raise Awareness

Educate users and stakeholders about the risks posed by deepfake manipulation and the importance of vigilant detection and mitigation strategies. By raising awareness about the prevalence of synthetic media manipulation and providing guidance on identifying and reporting suspicious content, you can empower individuals to protect themselves against malicious manipulation.

By implementing these best practices and adopting a proactive approach to deepfake detection, organizations can effectively safeguard against the spread of synthetic media manipulation and protect the integrity of digital content. Continued research, collaboration, and innovation are essential to stay ahead of evolving deepfake technology and mitigate the potential risks posed by malicious manipulation of media content.

Paravision Deepfake Detection

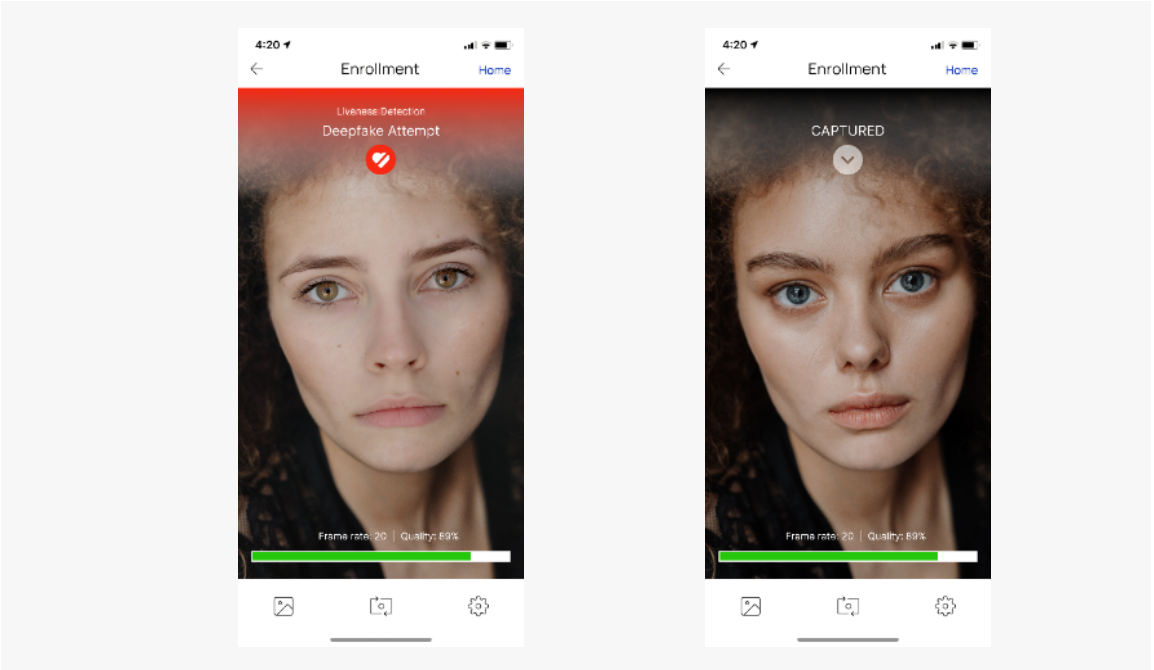
Paravision Deepfake Detection is one available tool against the rising tide of digital face manipulations in various sectors.

Leveraging AI-based analysis, Paravision Deepfake Detection assesses the likelihood of digital face manipulation using leading deepfake generation technologies, starting with Face Swaps (i.e., Identity Swaps) and Expression Swaps. By delivering an output score to guide automated or human-in-the-loop fraud analysis, organizations can quickly identify and mitigate the impact of deepfake threats.

Benefits of Paravision Deepfake Detection include:

- **Highly accurate:** AI-based analysis to assess the likelihood that still images or videos have been digitally manipulated with a broad range of leading deepfake generation technologies.
- **Ethically trained:** Based on a proprietary dataset of over 1 million deepfake images generated with properly-consented imagery and video, representing broad racial and gender diversity.
- **Easy to integrate:** Available as a part of Paravision's Docker container-based products, enabling cloud-ready solutions and simple integration for both new and existing Paravision partners.
- **Backed by extensive R&D:** Paravision Deepfake Detection results from years of focused R&D in collaboration with a Five Eyes Government partner.

With Paravision Deepfake Detection, organizations can confidently combat the threats posed by deepfake technologies, safeguarding identities, preserving trust, and upholding the integrity of digital content. If you would like to meet with our experts to hear more about Paravision Deepfake Detection, please contact us at info@paravision.ai.



In Conclusion

In the constantly shifting realm of synthetic media manipulation, the broad spectrum of deepfakes and digital face manipulations has emerged as a central concern, creating a need for robust detection to protect against the spread of misinformation and malicious content. This Practical Guide to Deepfakes underscores the importance of understanding these manipulations' diverse and sophisticated nature, which range from identity swaps to fully synthetic faces, each presenting unique challenges to digital security and authenticity.

AI plays a critical role in countering these challenges. Advanced AI-powered detection systems, which utilize machine learning models, pattern recognition, and multi-modal analysis, have been instrumental in identifying and curbing the proliferation of varied deepfake technologies across multiple platforms. These technologies are pivotal for recognizing traditional deepfakes and addressing the nuances of more complex manipulations that might not be classified strictly as deepfakes but are equally deceptive and harmful.

As we look to the future, the landscape of deepfake detection presents both hurdles and prospects. With deepfake technologies evolving rapidly, detection systems must adapt continuously to keep pace with new methods of digital deception. Ongoing research and development in AI-driven techniques, combined with robust industry collaborations and regulatory efforts, are essential for staying ahead of these threats. Furthermore, the ethical deployment of these technologies is paramount. By ensuring that AI models are ethically trained and utilize diverse datasets, detection systems can improve accuracy while ensuring fairness, reducing bias, and maintaining accountability.

The complexity of deepfakes and related digital face manipulations demands vigilant and innovative approaches to safeguard digital content. Through proactive measures, strategic collaborations, and adherence to high ethical standards, we can mitigate the risks associated with synthetic media. By staying informed and responsive to the ever-expanding variety of deepfakes, we can foster a more secure and trustworthy digital environment for all.

© Paravision 2024

PARAVISION

Trusted Identity AI

For more information or to schedule a demo, please contact us at:

info@paravision.ai

© Paravision 2024