Examining the Implications of Deepfakes for Election Integrity

Anonymous submission

Abstract

It is becoming cheaper to launch disinformation operations at scale using AI-generated content, in particular 'deepfake' technology. We have observed instances of deepfakes in political campaigns, where generated content is employed to both bolster the credibility of certain narratives (reinforcing outcomes) and manipulate public perception to the detriment of targeted candidates or causes (adversarial outcomes). We discuss the threats from deepfakes in politics, highlight model specifications underlying different types of deepfake generation methods, and contribute an accessible evaluation of the efficacy of existing detection methods. We provide this as a summary for lawmakers and civil society actors to understand how the technology may be applied in light of existing policies regulating its use. We highlight the limitations of existing detection mechanisms and discuss the areas where policies and regulations are required to address the challenges of deepfakes.

Introduction

In recent years, the digital world has undergone rapid advancements, resulting in the emergence of sophisticated technologies that blur the boundaries between reality and fiction. During these events, deepfake technology has been a cause for concern due to its potential implications and dangerous consequences. Coined from the terms "deep learning" and "fake," deepfakes utilize advanced algorithms to generate hyper-realistic multimedia content, often indistinguishable from authentic material (Mirsky and Lee 2021).

While AI-generated images and deepfakes both make use of artificial intelligence, they serve vastly different purposes primarily arising from the intent behind content creation and distribution. The former involves the creation of realistic images through algorithms that learn from real data, whereas the latter–technologically a subset thereof–typically aims to deceive viewers into believing they are accessing authentic content. This article (Becker and Laycock 2023) discusses the utilization of advanced technologies to produce lifelike and personalized dynamic facial visuals, as well as developing and adjusting various high-caliber static content. However, (Wang et al. 2022) underscores how intent matters, showcasing the potential abuse of deepfakes in producing fake images for scientific publications.

There are a variety of methods used to generate deepfakes; with state-of-the-art approaches including diffusion-based models and generative adversarial networks (GANs). Stable diffusion models, such as DALL-E 2, Midjourney, and Stable Diffusion are neural networks trained on a large dataset of images and captions to generate convincing images from text descriptions (Chen et al. 2023). On the other hand, GANs such as Cycle-GAN, DCGAN, and WGAN are deep learning systems commonly used for image generation, data augmentation, music generation, and deepfake creation (Remya Revi, Vidya, and Wilscy 2021). The quality of deepfakes generated using GANs depends on the quantity and variety of training data, and the use of GANs to synthesize minimum training data for deepfake generation has been an area of active research (Singh, Sharma, and Smeaton 2020).

The Dangers of Deepfakes for Democratic Elections

The use of deepfake technology to spread disinformation poses a significant threat to free and fair democratic elections. Deepfakes can serve as a potent tool for malicious actors to manipulate voters and influence election outcomes (McKenzie 2023; Appel and Prietzel 2022). There are several ways deepfakes endanger democratic processes:

1. Deepfakes can directly alter voter preferences and spread disinformation about candidates by making them appear to take policy positions they do not hold or engage in illegal behavior (Ray 2021; Appel and Prietzel 2022). This could undermine trust in the electoral process.
2. Coordinated disinformation campaigns utilizing deepfake videos could prevent citizens from voting by spreading false information about election procedures or intimidating voters through blackmail (Pawelec 2022). This form of voter suppression damages participation.
3. Deepfakes amplified through social media and messaging platforms can rapidly reach millions of viewers (Christopher 2020; Jee 2020), confusing them about candidates and issues. Widespread false or misleading information shaped by deepfakes harms informed civic discourse.
4. There are already instances of political deepfakes circulating globally, such as the manipulated videos of Nancy Pelosi and Volodymyr Zelenskyy (News 2019; Miller 2022). In the 2023 Argentina elections, both leading candidates, Javier Milei and Sergio Massa, created and

spread deepfake images and videos of each other to portray their opponent negatively (David Feliba 2023). This demonstrates the real-world vulnerability of elections.

DeepFake Creation and Identification

Generative Models

Deepfakes are fueled by technological advancements in generative models, broadly including autoencoders, GANs, transformer-based models, and diffusion-based models.

The historical developments in deepfake technology are as follows (Masood et al. 2023):

- **Pre-2014: Traditional Techniques and Early Autoencoders:** Before 2014, the field of manipulated multimedia predominantly utilized conventional methods such as splicing and copy-move, with an early occurrence dating as far back as 1860. Autoencoders, a generative model originating in the 1980s, garnered interest in the early 2000s and made significant contributions to the development of early generative models. Nevertheless, their influence on the progression of deepfake technology was diminished by more sophisticated models.
- **2014-2017: Emergence of GANs:** In 2014, the introduction of Generative Adversarial Networks (GANs) by Ian Goodfellow (Goodfellow et al. 2020) brought about a significant change in deepfake technology. During this time, GANs emerged as a highly influential and transformative factor, with academic initiatives such as Face2Face and Synthesizing Obama playing a significant role in the initial advancements. In September 2017, a significant event took place on Reddit when a user named "deepfake" shared the initial authentic deepfake. This entailed the utilization of computer-generated videos showcasing renowned actresses with their faces effectively replaced with explicit content. This occurrence garnered public interest, indicating a significant turning point in the advancement of deepfake technology. The event brought to light the fact that complex generative models could be used for bad and dishonest purposes, which raised awareness and led to more thought about ethical issues and regulatory actions.
- **2018 Onwards: Integration of Transformers and Diffusion Models:** Over the following years, the deepfake technology landscape continued to progress. The advent of open-source projects such as DeepFaceLab in 2018 has played a significant role in making deepfake creation tools more accessible. Furthermore, there was a significant change in the investigation of transformers beyond their original utilization in natural language processing. Researchers have acknowledged the adaptability of transformers, expanding their application to include image synthesis and other tasks unrelated to text.

Currently, the advancements in deepfake technology revolve around the integration of transformers and diffusion models. The objective of this collaborative approach is to attain generative outcomes of superior quality, with a focus on enhancing the authenticity and capabilities of the produced content. This technology advances with a focus on enhancing security. The progress is crucial, especially in the context of elections, where the threat of deepfakes contributing to misinformation campaigns continues to be a significant concern.

GAN-based Architectures

Very broadly, existing deepfake detection techniques can be divided into 2 categories, based on the consideration of change of characteristic/genuine attributes in space (spatial consideration) and the ones that consider changes in space as well as time (spatio-temporal). For the first class of techniques, researchers aim to capture spatial features and perform the classification of visual data (video/image) as real or fake (Bonettini et al. 2020; Raza, Munir, and Almutairi 2022; Ismail et al. 2021; Afchar et al. 2018; Coccomini et al. 2022a). On the other hand, models utilizing time-series and image features in combination have proven to be more accurate in identifying deepfakes (de Lima et al. 2020; Zhang et al. 2022; Coccomini et al. 2022b; Cai et al. 2023; Jung, Kim, and Kim 2020; Tariq, Lee, and Woo 2020; Cozzolino et al. 2021; Wodajo, Atnafu, and Akhtar 2023).

Evaluating GAN-based DeepFake Generation

Accuracy Ratings: Prominent tools like FaceSwap-GAN (An 2022) exhibit superior accuracy since perceptual loss improves the direction of eyeballs to be more realistic and consistent with the input face. It also smoothes out artifacts in the segmentation mask, resulting in higher output quality and rendering dependable options for a range of uses. Similarly, tools like Simswap, Fewshot FT Gan, and Faceshifter boast high accuracy. But it's important to recognize that some tools—like FaceApp and StyleGAN—have poorer accuracy levels. StyleGAN (Brownlee 2020) which relies on traditional GAN generators, inherits the interpretability and control issues associated with typical GAN models. Due to the generators' limited knowledge of latent space qualities and image synthesis techniques, it may be difficult to comprehend and manipulate components inside the StyleGAN framework.

Usability Analysis: CycleGAN(Zhu et al. 2017) stands out for its exceptional user-friendliness and versatility. Its simplicity lies in the elimination of the requirement for paired data (Paired training data consists of training examples x i, y i N i=1 , where the correspondence between x i and y i exists), making it accessible to a broader user base. The capability to seamlessly work with unpaired data not only simplifies the process but also proves cost-effective, addressing challenges associated with obtaining extensive and reliable paired datasets. This user-friendly approach, coupled with high accuracy in image-to-image translation, positions CycleGAN as a valuable tool. The usability of face swapping and attribute manipulation tools like Faceapp, SimSwap, Fewshot FT GAN, and FaceShifter depends on factors such as user interface, documentation, and the level of technical expertise required.

Security Assessment: By avoiding information loss, improving representation ability, and rejecting an attribute-independent constraint, AttGAN (He et al. 2019) prioritizes a secure facial attribute editing approach. The framework's security is strengthened by the attribute classification constraint on generated images, which guarantees accurate attribute manipulations. Utilizing adversarial learning and reconstruction provides additional resilience to maintain original facial features and produce realistic images. However, these implementation strategies have varied levels of defense against adversarial attacks, so they must be carefully considered. Overall, compared to models with more stringent constraints, AttGAN's combination of these features improves security

Computational Efficiency: CycleGAN is recognized for its computational efficiency, especially in handling unpaired data. Usability varies, with CycleGAN being user-friendly, while others, like Style-GAN variants, may demand deep learning expertise and substantial computational resources.

Challenges faced by existing tools and techniques

Challenges arise in the performance of deepfake detection algorithms when faced with low-quality films compared to high-resolution videos. Videos may undergo various transformations, including reshaping, rotations, and compression, necessitating flexible detection algorithms to maintain efficacy.

Time consumption emerged as a significant concern for real-world applications of deepfake-detection techniques. Despite their potential impact on social security, existing detection methods still face limitations in terms of extensive time requirements, hindering widespread adoption in practical scenarios.

The challenge of insufficient data for specific characters during the creation of deepfake models was highlighted. While models are often trained on specific datasets, they may struggle to produce accurate outputs when faced with limited data for a particular character. Retraining models for each distinct target character is a time-consuming process.

Dataset quality was identified as another challenging area, with most datasets created under ideal conditions that differ from real-world testing scenarios. This misalignment in dataset quality adds complexity to the development and evaluation of deepfake-detection algorithms.

Despite the availability of various deepfake-generation tools, inherent flaws and limitations persist. These tools are often tailored to specific traits, emphasizing the need for additional research to enhance their efficiency. Consequently, the creation of general-purpose deepfake-generation tools remains a complex and challenging process that warrants further investigation.

Reviewing Platform Policies against Deepfakes

The regulation of deepfake content on social media platforms has become a critical issue in recent years.

With the rise of AI-generated manipulated media, platforms like Meta, X (formerly Twitter), Reddit, Tiktok, and YouTube have implemented various policies to address the spread of deepfakes (Center 2023).

Meta

(Monika Bickert 2020)

- Removal of Manipulated Media: Meta will remove audio, photos, or videos, including deepfakes, if they violate any of their Community Standards, such as those related to graphic violence, voter suppression, and hate speech.
- Detection Efforts: Meta has launched the Deep Fake Detection Challenge and is collaborating with experts to address deepfakes and manipulated media.

X (formerly Twitter)

(X 2023)

- Prohibition of Misleading Media: X prohibits the sharing of synthetic, manipulated, or out-of-context media that may deceive or confuse people and lead to harm.

  However, memes, satire, animations, and cartoons are generally not in violation of this policy.
- Labeling and Consequences: In some cases, X may label posts containing misleading media and take action to reduce the visibility of the post on the platform.

YouTube

(PTI-News 2023)

- Disclosure Requirement: YouTube will require creators to disclose altered or synthetic content that is realistic, including using AI tools. The platform will inform viewers about such content through labels in the description panel and video player.
- Removal and Labeling: YouTube may remove AI-generated or manipulated content that simulates an identifiable individual, and it will work with creators to ensure they understand the new requirements.

Reddit

(Peters 2020)

- Reddit does not allow content that impersonates individuals or entities in a misleading or deceptive manner, including deepfakes or other manipulated content presented to mislead, or falsely attributed to an individual or entity.

Tiktok

(NBC-News 2023; Vincent 2023)

- TikTok bans deepfakes of private figures and young people, and all realistic AI deepfakes must be ”clearly disclosed”.

Monitoring the Implementation of Current Deepfake Regulations

Monitoring the enforcement of current deepfake regulations involves assessing the adherence of various stakeholders, evaluating the efficiency of detection tools, and addressing emerging challenges. We explore the strategies and considerations involved in overseeing the implementation of existing deepfake regulations:

1. **Stakeholder Compliance:** Regulatory success hinges on the compliance of key stakeholders, including social media platforms, technology companies, political campaigns, and content creators. Continuous monitoring of these entities is essential to ensuring they are actively adopting measures to prevent the creation and dissemination of malicious deepfakes. Collaborative efforts between regulatory bodies and stakeholders can facilitate the exchange of best practices, ensuring a unified approach to tackling deepfake-related threats.
2. **Effectiveness of Detection Tools:** The efficacy of deepfake detection tools plays a pivotal role in enforcing regulations. Regular assessments of the performance of existing detection mechanisms are crucial to identifying strengths, weaknesses, and areas for improvement. This involves evaluating the accuracy, speed, and adaptability of tools for detecting evolving deepfake techniques. Ongoing research and development are necessary to enhance the capabilities of detection tools and address emerging challenges in real time.
3. **Educational Initiatives:** Monitoring the implementation of regulations extends beyond enforcement measures to include educational initiatives. Informing the public, political candidates, and election officials about the existence of deepfake threats, the regulatory framework in place, and preventive measures is vital. Periodic assessments of the effectiveness of educational campaigns can guide adjustments and refinements to ensure they remain relevant and impactful.
4. **Adaptability to Evolving Threats:** The landscape of deepfake technology is dynamic, with new advancements and variations emerging regularly. Monitoring the implementation of regulations requires a proactive approach to stay ahead of evolving threats. Regulatory bodies should establish mechanisms for continuous threat assessment, allowing timely modification of regulations to address emerging challenges effectively.
5. **International Collaboration:** Given the global nature of information dissemination and potential cross-border impact, international collaboration is essential for effective regulation. Monitoring the implementation of deepfake regulations involves fostering partnerships between countries, sharing intelligence, and collectively addressing challenges. Regular forums for collaboration can facilitate the exchange of insights and strategies to combat the transnational aspects of deepfake threats.

Areas where interventions/policies and regulations are required

The emergence of deepfakes poses significant challenges to the integrity of elections. To address these challenges, various areas require targeted interventions, policies, and regulations.

1. **Authentication Protocols:** Implementing protocols to authenticate digital content can help distinguish genuine media from deepfakes. This includes the development of digital watermarks or certification systems (Westerlund 2019).
2. **Transparency Requirements:** Legislation mandating the disclosure of AI-manipulated content can increase transparency. Any altered media should be clearly labeled to inform the public about its modified nature (Langa 2021).
3. **Media Literacy Programs:** Educating the public, especially voters, about the existence and nature of deepfakes is crucial. Media literacy programs can teach people how to critically assess and verify the authenticity of the information they receive (El Mokadem 2023).
4. **Legal Frameworks Against Misuse:** There should be clear legal consequences for maliciously creating or distributing deepfakes (Feeney 2021).

Conclusion

This paper urgently calls for societal, policymaker, and regulatory action to safeguard elections from the pervasive threat of deepfake technology. It meticulously explores the dissemination of deepfakes and their implications for election security, advocating for comprehensive regulatory frameworks to counter the growing accessibility of AI-generated disinformation. The paper delves into the intricacies of the deepfake generation, emphasizing its profound impact on politics, trust, and democratic processes. While detection tools show promise, challenges persist, necessitating research and adaptability. Varied deepfake policies among technology and social media companies underscore the importance of vigilant monitoring for effective implementation. Recognizing the dynamic nature of these challenges, the paper urges continuous proactive efforts from policymakers, technologists, and the public to enhance detection capabilities and safeguard democracy against the evolving threats of deepfake technology. Ongoing commitment to innovation, collaboration, and democratic principles is crucial for ensuring the resilience of electoral processes against AI-generated disinformation. We also express our sincere gratitude to Mr. Eric Davis for his invaluable feedback and insightful comments on this paper.

References

Afchar, D.; Nozick, V.; Yamagishi, J.; and Echizen, I. 2018. MesoNet: a Compact Facial Video Forgery Detection Network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE.

An, S. 2022. Shaoanlu/faceswap-gan: A denoising autoencoder + adversarial losses and attention mechanisms for face swapping.

Appel, M.; and Prietzel, F. 2022. The detection of political deepfakes. *Journal of Computer-Mediated Communication*, 27(4): zmac008.

Becker, C.; and Laycock, R. 2023. Embracing deepfakes and AI-generated images in neuroscience research. *European Journal of Neuroscience*.

Bonettini, N.; Cannas, E. D.; Mandelli, S.; Bondi, L.; Bestagini, P.; and Tubaro, S. 2020. Video Face Manipulation Detection Through Ensemble of CNNs. arXiv:2004.07676.

Brownlee, J. 2020. A gentle introduction to stylegan the style generative Adversarial Network.

Cai, Z.; Ghosh, S.; Stefanov, K.; Dhall, A.; Cai, J.; Rezatofighi, H.; Haffari, R.; and Hayat, M. 2023. MARLIN: Masked Autoencoder for facial video Representation LearnINg. arXiv:2211.06627.

Center, S. U. D. 2023. Social Media Policies: Mis/Disinformation, Threats, and Harassment. https://statesuniteddemocracy.org/resources/social-media-policies/. [Online; accessed 30-November-2023].

Chen, Y.; Haldar, N. A. H.; Akhtar, N.; and Mian, A. 2023. Text-image guided Diffusion Model for generating Deepfake celebrity interactions. arXiv:2309.14751.

Christopher, N. 2020. We've Just Seen the First Use of Deepfakes in an Indian Election Campaign. https://www.vice.com/en/article/jgedjb/the-first-use-of-deepfakes-in-indian-election-by-bjp. [Online; accessed 21-November-2023].

Coccomini, D. A.; Messina, N.; Gennaro, C.; and Falchi, F. 2022a. *Combining EfficientNet and Vision Transformers for Video Deepfake Detection*, 219–229. Springer International Publishing. ISBN 9783031064333.

Coccomini, D. A.; Zilos, G. K.; Amato, G.; Caldelli, R.; Falchi, F.; Papadopoulos, S.; and Gennaro, C. 2022b. MINTIME: Multi-Identity Size-Invariant Video Deepfake Detection. arXiv:2211.10996.

Cozzolino, D.; Rössler, A.; Thies, J.; Nießner, M.; and Verdoliva, L. 2021. ID-Reveal: Identity-aware DeepFake Video Detection. arXiv:2012.02512.

David Feliba, T. R. F. 2023. How Argentina's new president-elect used AI to target his opponents during the campaign.

de Lima, O.; Franklin, S.; Basu, S.; Karwoski, B.; and George, A. 2020. Deepfake Detection using Spatiotemporal Convolutional Networks. arXiv:2006.14749.

DeepTraceTechnologies. 2023. "https://www.deeptracetech.com/. [Online; accessed 25-November-2023].

El Mokadem, S. S. 2023. The Effect of Media Literacy on Misinformation and Deep Fake Video Detection. *Arab Media & Society*, (35).

Feeney, M. 2021. Deepfake Laws Risk Creating More Problems Than They Solve. *Regulatory Transparency Project*.

Goldstein, J. A.; and DiResta, R. 2022. Research Note: This Salesperson Does Not Exist: How Tactics from Political Influence Operations on Social Media are Deployed for Commercial Lead Generation. *Harvard Kennedy School Misinformation Review*, 3(5): 1–15.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Hasan, H. R.; and Salah, K. 2019. Combating deepfake videos using blockchain and smart contracts. *Ieee Access*, 7: 41596–41606.

He, Z.; Zuo, W.; Kan, M.; Shan, S.; and Chen, X. 2019. Attgan: Facial attribute editing by only changing what you want. *IEEE transactions on image processing*, 28(11): 5464–5478.

Ismail, A.; Elpeltagy, M.; S. Zaki, M.; and Eldahshan, K. 2021. A New Deep Learning-Based Methodology for Video Deepfake Detection Using XGBoost. *Sensors*, 21(16).

Jee, C. 2020. An Indian politician is using deepfake technology to win new voters. https://www.technologyreview.com/2020/02/19/868173/an-indian-politician-is-using-deepfakes-to-try-and-win-voters/. [Online; accessed 30-November-2023].

Jung, T.; Kim, S.; and Kim, K. 2020. DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern. *IEEE Access*, 8: 83144–83154.

Langa, J. 2021. Deepfakes, real consequences: Crafting legislation to combat threats posed by deepfakes. *BUL Rev.*, 101: 761.

Masood, M.; Nawaz, M.; Malik, K. M.; Javed, A.; Irtaza, A.; and Malik, H. 2023. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied intelligence*, 53(4): 3974–4026.

McKenzie, B. 2023. IS THAT REAL? DEEPFAKES COULD POSE DANGER TO FREE ELECTIONS. https://news.virginia.edu/content/real-deepfakes-could-pose-danger-free-elections. [Online; accessed 20-November-2023].

Miller, J. R. 2022. Deepfake video of Zelensky telling Ukrainians to surrender removed from social platforms. https://nypost.com/2022/03/17/deepfake-video-shows-volodymyr-zelensky-telling-ukrainians-to-surrender/. [Online; accessed 21-November-2023].

Mirsky, Y.; and Lee, W. 2021. The Creation and Detection of Deepfakes: A Survey. *ACM Comput. Surv.*, 54(1).

Monika Bickert, G. P. M., Vice President. 2020. Enforcing Against Manipulated Media. https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/. [Online; accessed 30-November-2023].

NBC-News. 2023. TikTok bans deepfakes of young people as it updates guidelines. https://www.nbcnews.com/tech/tech-news/tiktok-bans-deepfakes-young-people-updates-guidelines-rcna75949. [Online; accessed 30-November-2023].

News, C. 2019. Doctored Nancy Pelosi video highlights threat of "deepfake" tech. https://www.cbsnews.com/news/doctored-nancy-pelosi-video-highlights-threat-of-deepfake-tech-2019-05-25/. [Online; accessed 21-November-2023].

Pawelec, M. 2022. Deepfakes and democracy (theory): how synthetic audio-visual media for disinformation and hate speech threaten core democratic functions. *Digital society*, 1(2): 19.

Peters, J. 2020. Reddit bans impersonation on its platform. https://www.theverge.com/2020/1/9/21058803/reddit-account-ban-impersonation-policy-deepfakes-satire-rules. [Online; accessed 30-November-2023].

PTI-News. 2023. YouTube Users Have To Disclose Altered Content That Looks Realistic: Google.

https://www.bqprime.com/nation/youtube-users-have-to-disclose-altered-content-that-looks-realistic-google. [Online; accessed 30-November-2023].

Ray, A. 2021. *The University of New South Wales Law Journal*, 44(3): 983–1013.

Raza, A.; Munir, K.; and Almutairi, M. 2022. A Novel Deep Learning Approach for Deepfake Image Detection. *Applied Sciences*, 12(19).

Remya Revi, K.; Vidya, K.; and Wilscy, M. 2021. Detection of Deepfake Images Created Using Generative Adversarial Networks: A Review. In *Second International Conference on Networks and Advances in Computational Technologies: NetACT 19*, 25–35. Springer.

SensityAI. 2023. https://sensity.ai/. [Online; accessed 25-November-2023].

Singh, S.; Sharma, R.; and Smeaton, A. F. 2020. Using GANs to Synthesise Minimum Training Data for Deepfake Generation. arXiv:2011.05421.

Tariq, S.; Lee, S.; and Woo, S. S. 2020. A Convolutional LSTM based Residual Network for Deepfake Video Detection. arXiv:2009.07480.

Vincent, J. 2023. TikTok bans deepfakes of nonpublic figures and fake endorsements in rule refresh. https://www.theverge.com/2023/3/21/23648099/tiktok-content-moderation-rules-deepfakes-ai. [Online; accessed 30-November-2023].

Wang, L.; Zhou, L.; Yang, W.; and Yu, R. 2022. Deepfakes: a new threat to image fabrication in scientific publications? *Patterns*, 3(5).

Westerlund, M. 2019. The emergence of deepfake technology: A review. *Technology innovation management review*, 9(11).

Wodajo, D.; Atnafu, S.; and Akhtar, Z. 2023. Deepfake Video Detection Using Generative Convolutional Vision Transformer. arXiv:2307.07036.

X. 2023. Synthetic and manipulated media policy. https://help.twitter.com/en/rules-and-policies/manipulated-media. [Online; accessed 30-November-2023].

Zhang, D.; Lin, F.; Hua, Y.; Wang, P.; Zeng, D.; and Ge, S. 2022. Deepfake Video Detection with Spatiotemporal Dropout Transformer. arXiv:2207.06612.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.

Appendix

From Theory to Practice: Deepfake Detection Examples

Analyzing the effectiveness of deepfake-detection tools.

1. **Sensity AI**(SensityAI 2023): Sensity AI is essential for identifying GAN-generated images, especially those that belong to the StyleGAN2 model. Sensity's model was used to analyze 975 profile pictures from LinkedIn accounts. The research study focused on fictitious accounts on the platform with GAN-generated profile photos (Goldstein and DiResta 2022). As evidenced by the results, there was over 90% confidence in the ability to identify GAN-generated images for 968 profile pictures. Surprisingly, 900 of these images had a confidence score higher than 99.9%, demonstrating the resilience of Sensity's model. The majority of the identified images were linked to the StyleGAN2 model, highlighting the efficacy of Sensity AI in detecting specific types of GAN-generated content.

2. **Truepic**: A startup company based in the United States has created a system that utilizes mobile apps to allow regular users and freelancers to capture images and store them on the company's servers. The purpose of saving the images is to maintain their integrity. As a result, comparing any fraudulent attempt with the image kept on the servers makes it easy to spot. Truepic utilizes blockchain technology (Hasan and Salah 2019) to securely store metadata associated with saved images, guaranteeing their immutability. This method is highly dependent on placing a significant amount of trust in Truepic about the authenticity and integrity of the uploaded images. The operational details of incorporating logos, text tickers, subtitles, or closed captions into images or video frames are not readily apparent.

3. **Deeptrace**(DeepTraceTechnologies 2023): Deeptrace employs machine learning algorithms for detecting deepfake videos, showcasing high accuracy rates achieved through comprehensive audio-based, visual-based, and text-based analysis. Its real-time detection capabilities, scalability for analyzing large datasets, and continuous learning to adapt to new deepfake techniques position it as a versatile solution. Deeptrace effectively addresses the challenges posed by deepfakes, making it a valuable asset for organizations seeking reliable detection tools.