If You Give a Mouse a Vulnerability:
How the History of Malware Informs the Future of DeepFakes

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Kelly Schaefer

Spring, 2022

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

David Evans, Department of Computer Science

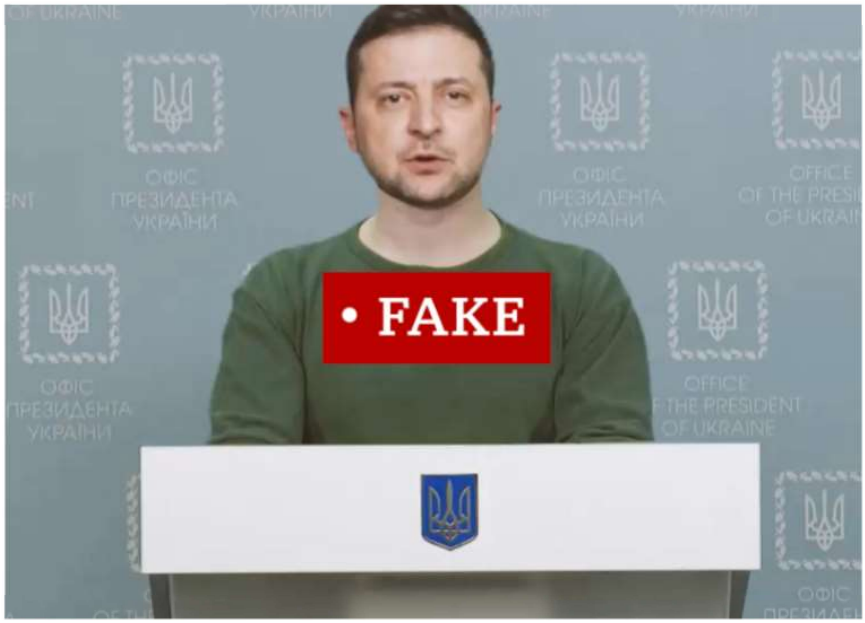
Technical Report

Abstract

On Wednesday March 16th, 2022, a group of hackers broadcast a fake video of Ukrainian President Volodymyr Zelenskyy informing soldiers to lay down their weapons and surrender to Russian forces (Allyn, 2022). The application of deep-learning based forgery to political manipulation exemplifies a threat of this type of content manipulation, known as DeepFakes, that has raised concerns since applications of deep-learning based face-swaps became popularized in 2017 (Aubé, 2017). The quality of the Zelenskyy DeepFake was not state of the art.It contained visual and auditory artifacts that allowed users to easily identify the video as fake. Most prominently, viewers referenced the inaccurate accent of the video audio. However, more sophisticated DeepFakes are not easily distinguishable from authentic content. In order to engage in academic discourse on the contemporary threat of DeepFakes, this technical paper overviews current DeepFake generation and detection methods, elucidates countermeasures, and summarizes the current performance of generation and detection. Once the technical background in contemporary DeepFake technology has been established, the paper draws on parallels to the race between malware generation and malware detection to inform technical predictions around the future trajectory of DeepFake generation and mitigation.

Figure 1

Labelled Screenshot of the Volodymyr Zelensky DeepFake Released on TV24's Website



Note. Image taken from Sardarizadeh, S. [@Shayan86]. (2022, March 16). *President Zelensky has uploaded a video to refute the fake video of him … TV24's hacked website still has a screenshot of the fake video along with a transcript of it. [Tweet].* Twitter.

https://twitter.com/Shayan86/status/1504106312115888130?s=20&t=Q-78n2FiwZDZA1t6Iv6EjQ

1. DeepFake Generation Technologies

The computing research community has investigated applications of machine learning to manipulate visual media for decades. In August 1997, a group of researchers presented on the application of computer vision to learn and replicate the visual speech patterns of a particular subject. This research was presented as a method to improve film dubbing by syncing lip motion to new audio-tracks (Bregler et. al., 1997). Since then, computer vision methodologies and capabilities have seen rapid advancements. Correspondingly, the ability to convincingly fabricate and manipulate visual data has caught the attention of multifarious individuals with a wide range of intentions.

Figure 2

Deep Learning Based Pose Frontalization to Aid Facial Recognition

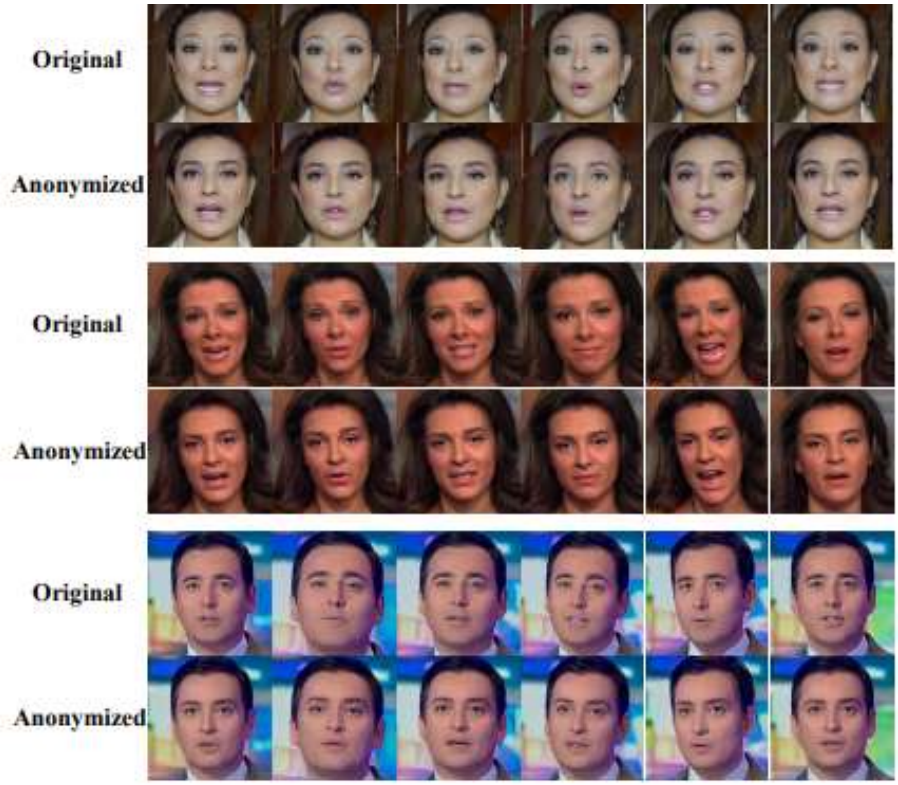
Note. Face images from CFP dataset and the synthesis images frontalized by the proposed GSP-GAN. Luan, X., Geng, H., Liu, L., Li, W., Zhao, Y., & Ren, M. (2020). Geometry Structure Preserving Based GAN for Multi-Pose Face Frontalization and Recognition. *IEEE Access*, *8*, 104676–104687. https://doi.org/10.1109/ACCESS.2020.2996637

Methodologies that use footage of a source subject to drive synthetic expressions on a target subject have been used to facilitate post-production in the movie and video game industries (Perov et. al, 2021). Advanced facial recognition systems use deep learning and

computer vision to change the pose of subjects in security footage as shown in Figure 2 (Luan et al., 2020). Additionally, displayed in Figure 3, facial swapping via computer vision is proposed as a technique to anonymize publicly available photographs and video (Ma et al., 2021; Rothkopf, 2020). However, the fabrication of media through deep learning algorithms gained widespread attention in 2017 when a Reddit user with the online pseudonym “DeepFakes” began proliferating pornographic videos manipulated to feature the faces of popular celebrities in the place of adult performers (Tolosana et. al., 2020). As a result, the term “DeepFake" has become synonymous with manipulated media generated via deep-learning. The threat of malicious DeepFake applications has garnered significant attention.

Figure 3

FaceSwap Based Identity Anonymization



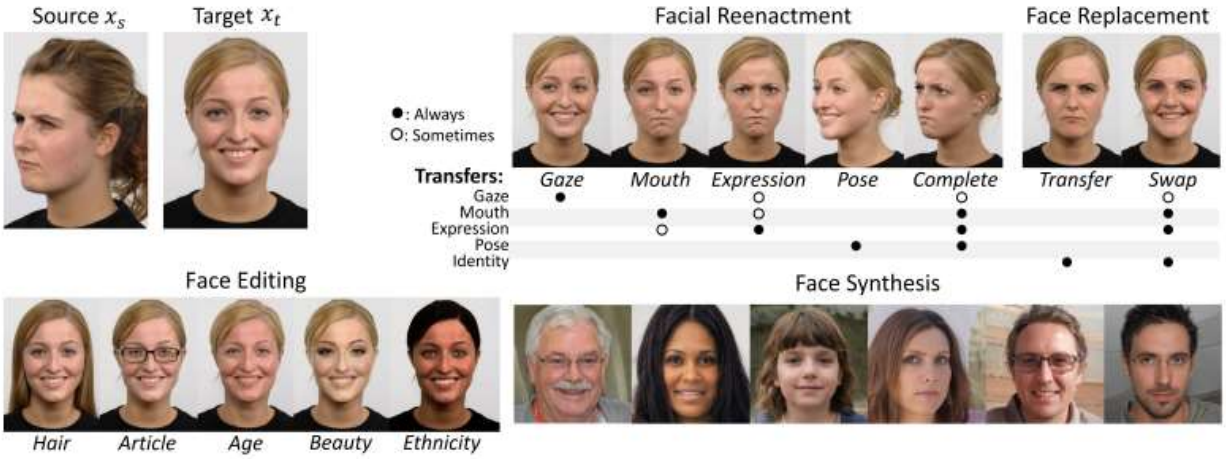
Note. Face anonymization effects of [Ma et al.]'s method on FaceForensics++ videos. The original face video frames and the corresponding anonymized faces are shown.Taken from Ma, T., Li, D., Wang, W., & Dong, J. (2021). CFA-Net: Controllable Face Anonymization Network with Identity Representation Manipulation. *ArXiv:2105.11137 [Cs]*. http://arxiv.org/abs/2105.11137

1.1 Kinds of DeepFakes

Literature on DeepFakes often classifies deep-learning manipulations into a few different categories. Common categories of visual DeepFakes are face-swap, reenactment, lip-syncing, face synthesis and attribute manipulation, though the names and boundaries between categories vary (Mirsky & Lee, 2022; Masood et al., 2021). Face-swaps have gained significant attention and are currently the most prevalent form of deepfake manipulation (Masood et. al, 2021).

Figure 4

Examples of Different Kinds of DeepFakes



Note. Examples of reenactment, replacement, editing, and synthesis deepfakes of the human face.Taken from Mirsky, Y., & Lee, W. (2022). The Creation and Detection of Deepfakes: A Survey. *ACM Computing Surveys*, *54*(1), 1–41. https://doi.org/10.1145/3425780

Face-swap DeepFakes replace the face of a subject in a source video with the identity of a target subject.

Reenactment DeepFakes use footage of a source subject to drive the expressions, pose, gaze or body movements of a target subject.

Lip-sync DeepFakes drive the motion of a subject's mouth to match a desired audio.

Attribute manipulations do not use a target identity, but rather manipulate features on a source subject. Manipulated features can include age, facial hair, clothing, weight, ethnicity, and beauty.

Face synthesis involves creating synthetic, realistic looking human faces.

Additionally, deep-learning based audio-manipulations are a category of DeepFake media that has seen recent growth in the wild (Masood et. al, 2021). Two primary approaches for generating audio DeepFakes are text-to-speech synthesis and voice conversion.

1.2 DeepFake Malicious Use

Malicious use of DeepFakes undermines trust in truth. Puppetry and face swap methods can be used to target individuals in defamation and discredibility attacks. A 2018 political defamation attack on journalist Rana Ayyub leveraged pornographic DeepFake technology to defame and discredit the young journalist by face-swapping her features onto an adult performer. This forged video was shared thousands of times, resulting in a level of cyber harassment that warranted intervention by the United Nations (Ayyub, 2018).The application of face-swap and reenactment videos to target political leaders in disinformation attacks has gained widespread attention. In May 2018, a political group in Belgium released a DeepFake video of Donald Trump urging Belgium to withdraw from the Paris Climate Agreement. (Schwartz, 2018). Later, in 2020, a climate activist group produced a forged lip-sync video of Belgian Prime Minister Shophie Wilmès speaking on the climate crisis (Galindo, 2020). Allegedly, both videos, shown in Figure 5 and 6 respectively, were produced with the intent to gain attention rather than deceive viewers. However, as mentioned in the introduction, Ukrainian President Volodymyr Zelenskyy was recently targeted in an attack that involved a face-swap video broadcast on live television. In this forgery, the Ukrainian President was depicted directing soldiers to lay down their weapons in the war against Russia (Allyn, 2022).

Figure 5

DeepFake Video of Donald Trump Urging Belgium to Withdraw From the Paris Climate Agreement

Note. Taken from Vooruit. (2018, May 20). *Teken de Klimaatpetitie Trump heeft een boodschap voor alle Belgen……* [Post]. Facebook. https://www.facebook.com/Vlaamse.socialisten/videos/10155618434657151/

Figure 6

DeepFake Video of Shophie Wilmès Speaking on the Climate Crisis

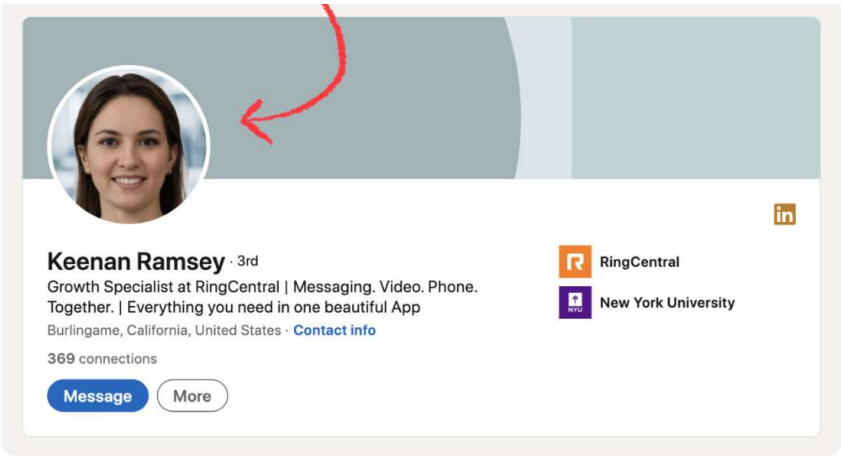
Note. Taken from Ajder, H. (2020, April 20). *Tracer newsletter #50 (20/04/20)-extinction rebellion release deepfake of Belgian prime minister...* Medium. Retrieved April 29, 2022, from https://medium.com/sensity/tracer-newsletter-50-20-04-20-extinction-rebellion-release-deepfake-of-belgian-prime-minister-2b48d586b44

In response to disinformation threats, legislation passed in Texas and California ban the distribution of deceptive audio or visual media around elections (AB 730 CA., 2019; SB 751 TX., 2019). Other states, such as Maine, Washington and Maryland, have introduced similar legislation (HB 198 MD, 2019; SB 1988 ME, 2020; SB 6513 WA, 2020). Legal courts also grapple with DeepFake threats. Malicious actors may employ face-swap and reenactment applications to tamper with visual evidence. This amplifies a threat, referred to as the liar's dividend, where any footage can be refuted as fake. In this way, visual evidence is rendered less reliable (Chesney & Citron, 2019).

Facial editing and synthesis are leveraged in online deception. Mirsky & Lee describe baiting by child predators as a malicious surface for deep learning based facial editing; predators can edit photos to appear younger (Mirsky & Lee, 2022). Profiles are also manipulated through facial synthesis models.Forged photos that resemble real people, exemplified in Figure 8, have been used in online profiles to spread disinformation and conduct corporate scams (Bond, 2022). Additionally, the European Union identifies GAN based face morphing as a method of creating falsified IDs that match the identity of two individuals (Ciancaglin et al., 2020).

Figure 8

DeepFake Profile Picture Created Via Facial Synthesis



Note. Taken from Bond, S. (2022, March 27). *That Smiling Linkedin Profile Face Might be a Computer-Generated Fake*. NPR. Retrieved April 29, 2022, from https://www.npr.org/2022/03/27/1088140809/fake-linkedin-profiles

In the wild, DeepFakes are increasingly seen in phishing schemes. Voice phishing attacks have applied auditory DeepFakes as a vector for impersonation (Brewster, 2021). Reenactment DeepFakes, such as pornographic face-swaps, have been used as a blackmail tool in other extortion schemes (Joshi, 2021).

1.3 Challenges to Realistic DeepFake Generation:

While research has worked to address limitations, existing models of DeepFake generation have certain weak points. Mirsky & Lee identify the following as challenges of creating realistic DeepFakes: generalization, paired training, identity leakage, occlusions, and temporal coherence.

Figure 9

DeepFakes with Low Fidelity Due to Challenging Generation Conditions

Note. Problems with current deepfake generation methods. From left to right: low resolution, low quality, strange artifacts due to wearable items, and facial pose variations.Taken from Le, T.-N., Nguyen, H. H., Yamagishi, J., & Echizen, I. (2022). *Robust Deepfake On Unrestricted Media: Generation And Detection*. doi:10.48550/ARXIV.2202.06228

Generalization refers to limitations of models to adapt to new identities, illumination conditions, head poses, and other image variations. Masood et al. note that pose variations, illumination conditions and distance from the camera can interfere with the production of quality DeepFakes (Masood et al., 2021). The best results are seen when the input media has a frontal facial view (Xuan et al., 2019). Varying illumination conditions between source and target images result in semantic inconsistencies in generated media.

Occlusion refers to challenges presented by objects that obscure the face, such as glasses or motioning hands. These occlusions can increase semantic inconsistencies in the generated output.

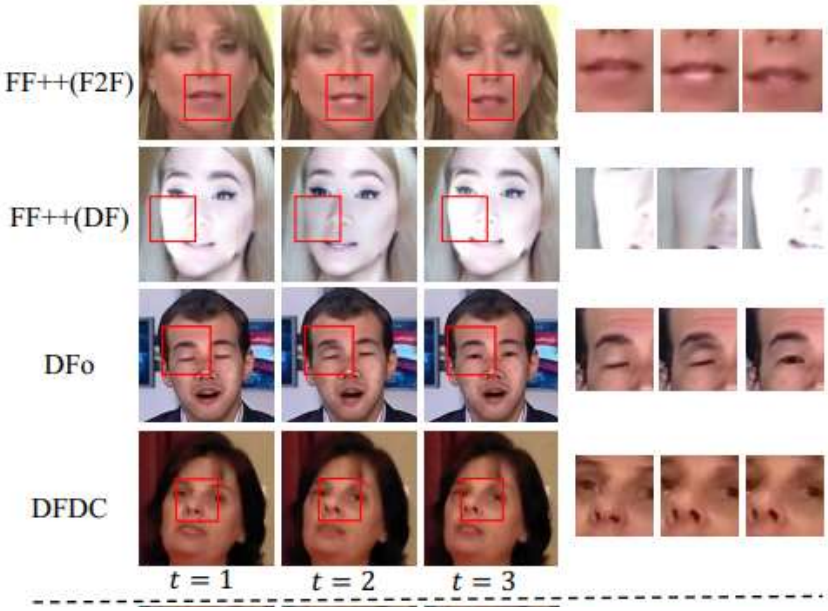
Identity leakage references weakness in face swap and reenactment DeepFakes where the source identity is reproduced along with the target identity in the generated DeepFake

Paired training refers to the need to match the input to a neural network with the desired output when training, which can be a laborious process.

. *Temporal coherence* refers to artifacts such as flicker, jitter, and semantic inconsistencies that appear when DeepFake generators operate on a frame by frame basis (Mirsky & Lee, 2022). Frame by frame temporal inconsistency is displayed in Figure 10.

Figure 10

Temporally Inconsistent Frames in DeepFake Datasets



Note. Temporal incoherence in existing datasets: FaceForensic++(FF++), DeeperForensics(DFo), Deepfake Detection Challenge Preview(DFDC), and FaceShifter(FSh). In the top 4 rows, we show four temporal incoherence that happened between the neighborhood frames Taken from Zheng, Y., Bao, J., Chen, D., Zeng, M., & Wen, F. (2021, Οκτώβριος). Exploring Temporal Coherence for More General Video Face Forgery Detection. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 15044–15054.

Additionally, current generation methods for synthetic audio have weakness in lack of natural emotions, lack of natural pauses, and behavioral variation from the target identity. Behavioral variations are observed in speaking pace and breathiness (Masood et al., 2021).

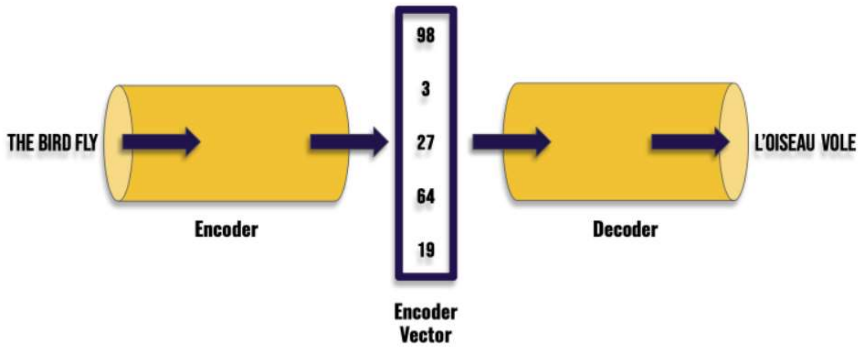
1.4 DeepFake Model Architecture

Encoder-Decoder Networks:

Early face swap models were built using two encoder-decoder network pairs (deepfakes, 2022/2017; Masood et al., 2021). Often used in translation applications, an encoder-decoder network, diagrammed in Figure 11, functions by training two separate networks on interpretation tasks. The first network, the encoder, takes in input data and transforms it into a statistical vector representation. The second network, the decoder network, takes the statistical vector and reproduces the input data (Keldenich, 2021). The application of encoder-decoder networks to face swap applications involves training the first encoder-decoder network on the source face and the second encoder-decoder network on the target face. The encoder extracts latent features and the decoder reconstructs the face. Then, the decoders are swapped. This results in a model that uses the source encoder and the target decoder to produce an image with the identity of the source face on the target image (Mirsky & Lee, 2022). An encoder-decoder network that learns without labels is known as an autoencoder (Kana, 2020).

Figure 11

Diagram of Encoder-Decoder Network



Note. Taken from Keldenich, T. (2021, October 17). Encoder Decoder What and Why ? - Simple Explanation. *Inside Machine Learning*. https://inside-machinelearning.com/en/encoder-decoder-what-and-why-simple-explanation/

Variational Autoencoder Networks:

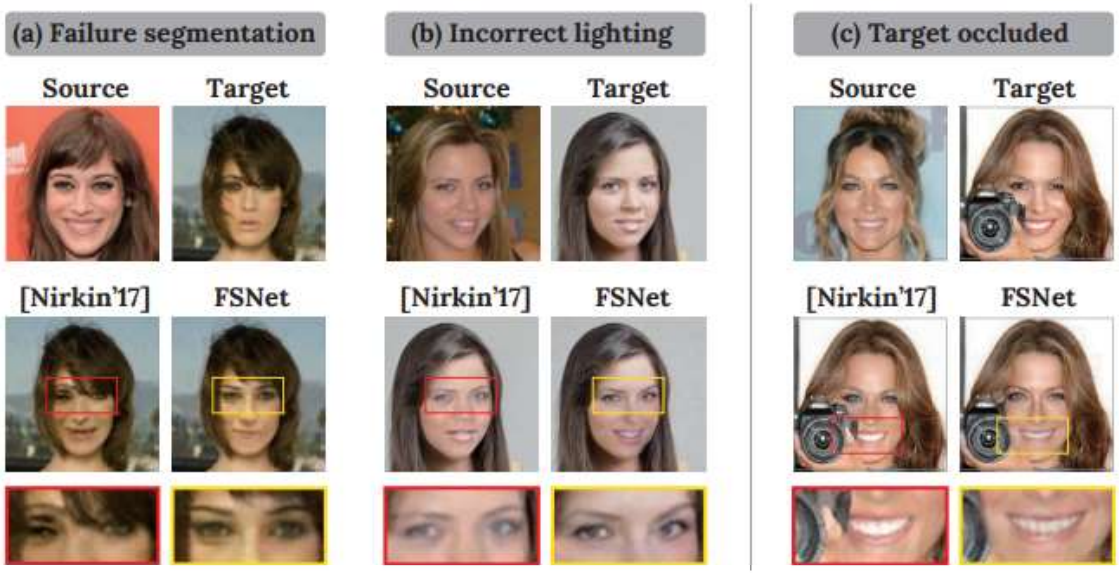
One weakness of standard autoencoder networks is a sparsely populated latent space. A research advance to improve the generation of images from this sparsely populated space is the introduction of normalization. By enforcing a normal distribution on the latent space, values are more continuous. A variational autoencoder computes the mean and standard deviation of latent

space values. Then, values sampled from a normal distribution are propagated to the decoder (Kana, 2020).

The application of variational autoencoder networks to DeepFake generation is often used in disentanglement to target training towards specific features. This disentanglement improves the image fidelity of DeepFakes. Additionally, variational autoencoder architectures improve the performance of DeepFake generation for source data with inappropriate fitting for 3D morphable models, as shown in Figure 12. These cases include strange lighting conditions and different facial orientations. However, a weakness of the variational autoencoder generation process is the loss of target lighting conditions and occlusions, such as glasses or hands (Masood et al., 2021). Table 1 surveys DeepFake models that leverage variational autoencoder networks.

Figure 12

Comparison of Face-Swaps using 3D Morphable Model and Variational Autoencoder Architecture Across Suboptimal Cases



Note. Typical failure cases of Nirkin et al.'s method [15] and FSNet. (a) Failure segmentation and (b) incorrect lighting estimation are those for Nirkin et al.'s method and (c) occluded target face is that for proposed FSNet. Taken from Natsume, R., Yatagawa, T., & Morishima, S. (2018). FSNet: An Identity-Aware Generative Model for Image-based Face Swapping. *ArXiv:1811.12666 [Cs].* http://arxiv.org/abs/1811.12666

Table 1

DeepFake Generation with Variational Autoencoder Architecture

Model Name	**Authors**	**Description**	**Kinds of DeepFake Applications**
RSGAN	(Natsume et al., 2018a)	Region-separative generative adversarial network; utilizes 2 variational autoencoders that target training towards facial and hair regions separately	face-swap; attribute manipulation; synthesis
FSNET	(Natsume et al., 2018b)	Both variational autoencoder objectives and generative adversarial network objectives are used; trains toward face region in source images and non-face regions in target images; uses inpainting in the generator	face-swap
CVAE_GAN	(Bao et al., 2017)	Utilizes a combined variational autoencoder and generative adversarial network that conditions generation on fine grained categories	synthesis; attribute manipulation; inpainting*
Additive Focal Variational Auto-encoder	(Qian et al., 2019)	Targets training towards appearance encodings and identity-agnostic expression encodings	attribute manipulation; reenactment
LumièreNet	(Kim & Ganapathi, 2019)	Uses source audio to drive facial expressions, body postures and gestures	lip-sync

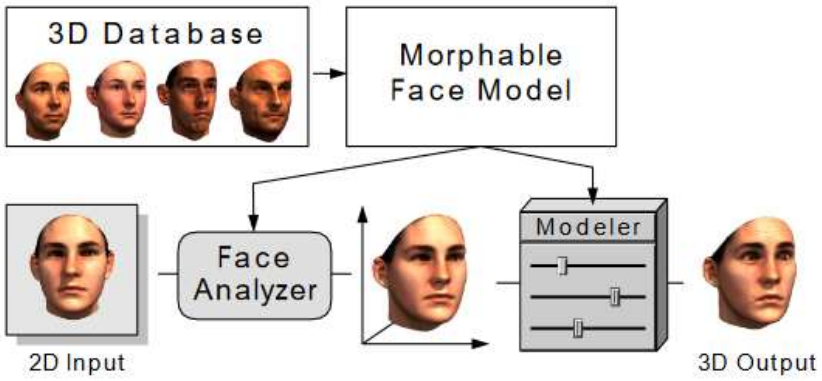
3D Morphable Models:

In 1999, Blanz and Vetter proposed a 3D morphable face model, depicted in Figure 13, which generates subject-specific 3D face models from 2D photographs. This goal is

accomplished through a substantial face model database that stores facial texture and shape as a vector representation. 3D faces are generated through the formation of linear combinations of prototype faces (Blanz & Vetter, 1999). 3D morphable models have recently seen a rise in deep learning applications (Egger et al., 2020). DeepFake generation, particularly reenactment generation, has leveraged these 3D morphable models. Table 2 overviews a variety of DeepFake architectures which apply 3D morphable models.

Figure 13

Diagram of Original 3D Morphable Face Model



Note. Derived from a dataset of prototypical 3D scans of faces, the morphable face model contributes to two main steps in face manipulation: (1) deriving a 3D face model from a novel image, and (2) modifying shape and texture in a natural way. Taken from Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '99*, 187–194. https://doi.org/10.1145/311535.311556

Table 2

DeepFake Generation with 3D Morphable Models

Model Name	**Authors**	**Description**	**Kinds of DeepFake Application**
Deep Video Portraits	(Kim et al., 2018)	Drives full head animation through a space-time generative adversarial network	reenactment; lip-sync; expression reenactment

FaceID-GAN	(Shen et al., 2018a)	Uses a 3-player generative adversarial network architecture in combination with a 3D morphable model to produce an architecture for facial reenactment with reduced identity leakage	reenactment
FaceFeat-GAN	(Shen et al., 2018b)	Uses three encoder-predictor networks in a 3-player generative adversarial network architecture. One of the encoder-predictor networks is trained to predict 3D morphable model parameters	reenactment; face swap
paGAN	(Nagano et al, 2018)	Utilizes a conditional generative adversarial network and 3D morphable model to produce 3D avatars from a single source image	reenactment
*n/a**	(Nirkin et al., 2017)	Uses a fully convolutional neural network to segment facial regions; applies a 3D morphable model to understand facial texture and geometry	face-swap
VDub	(Garrido et al., 2015)	Uses audio analysis and space-time retrieval in combination with a 3D morphable model to reproduce a mouth region that matches source audio	lip-sync

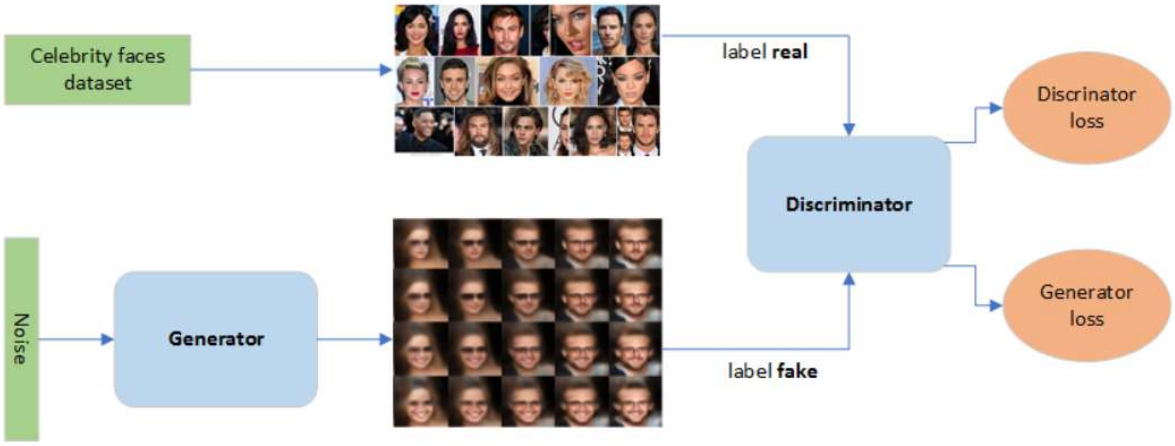
Generative Adversarial Networks:

Drawing on game theory foundations, the first generative adversarial network (GAN) was proposed in 2014 by Goodfellow et al (Goodfellow et al, 2014). This architecture functions by using an adversarial network to pit a generative model against a discriminative model. The generative model attempts to produce data matching the distribution of the training set and the discriminative model attempts to classify whether or not data was produced by the generative

model. The generator is trained towards maximizing the classification error of the discriminator. The discriminator is trained toward minimizing classification error (Goodfellow et. al, 2014; Kana, 2021). Generative adversarial networks have become prominent in the generation of DeepFakes. Table 3 surveys DeepFake generation methods that leverage generative adversarial network architecture.

Figure 14

Diagram of Generative Adversarial Network Using a Celebrity Faceset



Note. Taken from Kana, M. (2021, February 19). *Generative Adversarial Network (GAN) for Dummies—A Step By Step Tutorial.* Medium. https://towardsdatascience.com/generative-adversarial-network-gan-for-dummies-a-step-by-step-tutorial-fdefff170391

In 2017. Karras et al. improved generative adversarial network training through their ProGAN methodology, which utilizes a mini-batch size and additional network layers to increase the resolution of generated images (Karras et al., 2018). StyleGAN and StyleGAN2 build further on ProGAN to improve fidelity (Karras et al., 2019; Karras et al.; 2020). StyleGAN2 addresses semantic attributes (Karras et al., 2020). In the case of DeepFakes, these semantic attributes can include gaze direction and teeth alignment. A number of other researchers have used other varying strategies to address the resolution of GAN generated images (Zhang, Goodfellow, et al., 2019; Brock et al., 2019).

When using generative adversarial networks for DeepFake generation tasks, there is a necessity for a large corpus of high fidelity training data. Facial reenactment models that require less subject-specific training data address this weakness (Zakharov et al., 2019; Zhang, Zhang, He et al., 2019; Hao et al., 2020). However, weaknesses of these few-shot learning methods include identity leakage, where a source identity is partially reproduced in generated videos (Masood et al., 2021). DeepFake generation has also addressed temporal incoherence through the

use of temporal coherence analysis and optical flow estimation in discriminators (Masood et al., 2021).

Table 3

DeepFake Generation with Generative Adversarial Networks

Model Name	**Authors**	**Description**	**Kinds of DeepFake Application**
Faceswap-GAN	(shaoanlu, 2022/2017)	Adds a discriminator and adversarial loss to the encoder-decoder network popularized by reddit user deepfakes	faceswap
FSGAN	(Nirkin et al., 2019)	A subject agnostic model that uses a.) a network for face completion to handle weaknesses related to facial occlusions. and b.) a network for face blending to reduce artifacts, while maintaining target lighting	face-swap; reenactment
End-to-End Speech-Driven Realistic Facial Animation with Temporal GANs	(Vougioukas et al., 2019)	Utilizes discriminators to improve audio-visual synchronization in a model that uses audio to drive motions of a talking head	lip-sync
TP-GAN	(Huang, Zhang, et al., 2017)	Uses a two pathway generative adversarial network to handle minimal poorly posed input images.This architecture perceives global structure, through an encoder-decoder network, and local details through four patch networks. The architecture uses a combination loss function of adversarial loss, symmetry loss, and identity-preserving loss	reenactment
Deep Video Portraits	(Kim et al., 2018)	Drives full head animation through a space-time generative adversarial network	reenactment; lip-sync; expression reenactment

FaceID-GAN	(Shen et al., 2018a)	Uses a 3-player generative adversarial network architecture in combination with a 3D morphable model to produce an architecture for facial reenactment with reduced identity leakage	reenactment
FaceFeat-GAN	(Shen et al., 2018b)	Uses three encoder-predictor networks in a 3-player generative adversarial network architecture. One of the encoder-predictor networks is trained to predict 3D morphable model parameters	reenactment; face swap
paGAN	(Nagano et al, 2018)	Utilizes a conditional generative adversarial network and 3D morphable model to produce 3D avatars from a single source image	reenactment
RSGAN	(Natsume et al., 2018a)	Region-separative generative adversarial network; utilizes 2 variational autoencoders that target training towards facial and hair regions separately	face-swap; attribute manipulation; synthesis
FSNET	(Natsume et al., 2018b)	Both variational autoencoder objectives and generative adversarial network objectives are used; trains toward face region in source images and non-face regions in target images; uses inpainting in the generator	face-swap
CVAE_GAN	(Bao et al., 2017)	Utilizes a combined variational autoencoder and generative adversarial network that conditions generation on fine grained categories	synthesis; attribute manipulation; inpainting*
AttGAN	(He et al., 2018)	Applies attribute classification constraint to guarantee manipulation of desired features; applies reconstruction learning to constrain manipulation to only the desired features	attribute manipulation

PA-GAN	(He et al., 2020)	Applies an attention mask to constrain editing to an attribute area; progressively manipulates attributes from high level features to low level features	attribute manipulation
SaGAN	(Zhang et al., 2018)	Utilizes an attribute manipulation network and a spatial attention network in the generator of a generative adversarial network to restrict attribute manipulation to certain regions of an image	attribute manipulation
STGAN	(Liu et al., 2019)	Incorporates selective transfer units into an encoder-decoder network in the generator to improve accuracy and perceptual quality of attribute manipulations	attribute manipulation
StarGAN v2	(Choi et al., 2020)	Addresses the diversity and scalability of image to image translation across multiple domains using a generative adversarial network architecture with a mapping network and style encoder	attribute manipulation; reenactment; face swap
Pix2pixHD	(Wang et al., 2018)	Produces high resolution images through conditional adversarial networks with perceptual loss, a measure of high level differences in images	reenactment; attribute manipulation; face-swap; lip-sync
GANimation	(Pumarola et al., 2018)	Conditions generation on annotated Action Units (AU), which encode facial expressions; utilizes attention mechanisms to improve robustness to varying background and lighting conditions	reenactment
GAN with triple consistency loss	(Sanchez & Valstar, 2018)	Introduces a triple consistency loss to generative adversarial face translation to better handle differing distributions in the input and target domains	reenactment; face-swap

Few-Shot Adversarial Learning of Realistic Neural Talking Head Models	(Zakharov et al., 2019)	Extensively trains on a large dataset of videos to allow for the generation of realistic neural talking head models from one or few frame(s)	reenactment; face-swap
One-shot Face Reenactment	(Zhang, Zhang, He et al., 2019)	One shot learning approach, which disentangles shape and appearance information through 2 encoders with a shared decoder that aggregates multi-level features	reenactment
FaR-GAN	(Hao et al., 2020)	Uses a generative adversarial network which composes appearance and expression information for effective face modeling in one-shot reenactment	reenactment

General Advances in DeepFake Generation:

Recent notable advances in DeepFake generation include post-processing steps that address artifacts, address occlusions and improve smoothing. Inpainting, a computer vision task that fills in missing details in an image, improves the fidelity and coherence of generated DeepFakes. Additionally, artifacts are reduced through the use of loss functions that handle specific weaknesses (Masood et al., 2021). Perceptual loss based on the VGG-Face vision model is used to improve the fidelity of eye movements and to smooth artifacts (Masood et. al., 2021). Self-attention modules and adaptive instance normalization (AdaIN) layers improve image fidelity (Huang & Belongie, 2017; Masood et al, 2021). There have also been a variety of approaches to lower the burden of training data. Masood et al note advances in unpaired, self-supervised training strategies which mitigate a need for extensive labeled training data (Masood et al, 2021). Mirsky & Lee note variations among the generalizability of models. Rigid models require training toward a specific source identity and a specific target identity. More general models allow any source identity to drive the target identity that a model was trained on. The most generalizable models use any source identity to drive any target identity (Mirsky & Lee, 2022). Advances in DeepFake generation permit increased real time manipulations, allowing for a greater enmeshment with other social engineering pressures (Masood et al, 2021).

2. DeepFake Countermeasures

DeepFake generation is not flawless. As shown in Figure 15 and Figure 16, many DeepFakes contain visual, semantic artifacts, such as discoloration, inconsistent lighting, unnatural teeth, or unnatural hair that indicate inauthenticity to a viewer (Johansen, 2020). A variety of strategies have been proposed to mitigate the threat of DeepFake forgeries. Researchers have proposed a number of different deep learning based technical detection methods. Scholarship has introduced frameworks for digital providence. Some scholars have asserted a need for increased digital literacy. Other research has investigated technical adversarial attacks of DeepFake generation through image perturbations.

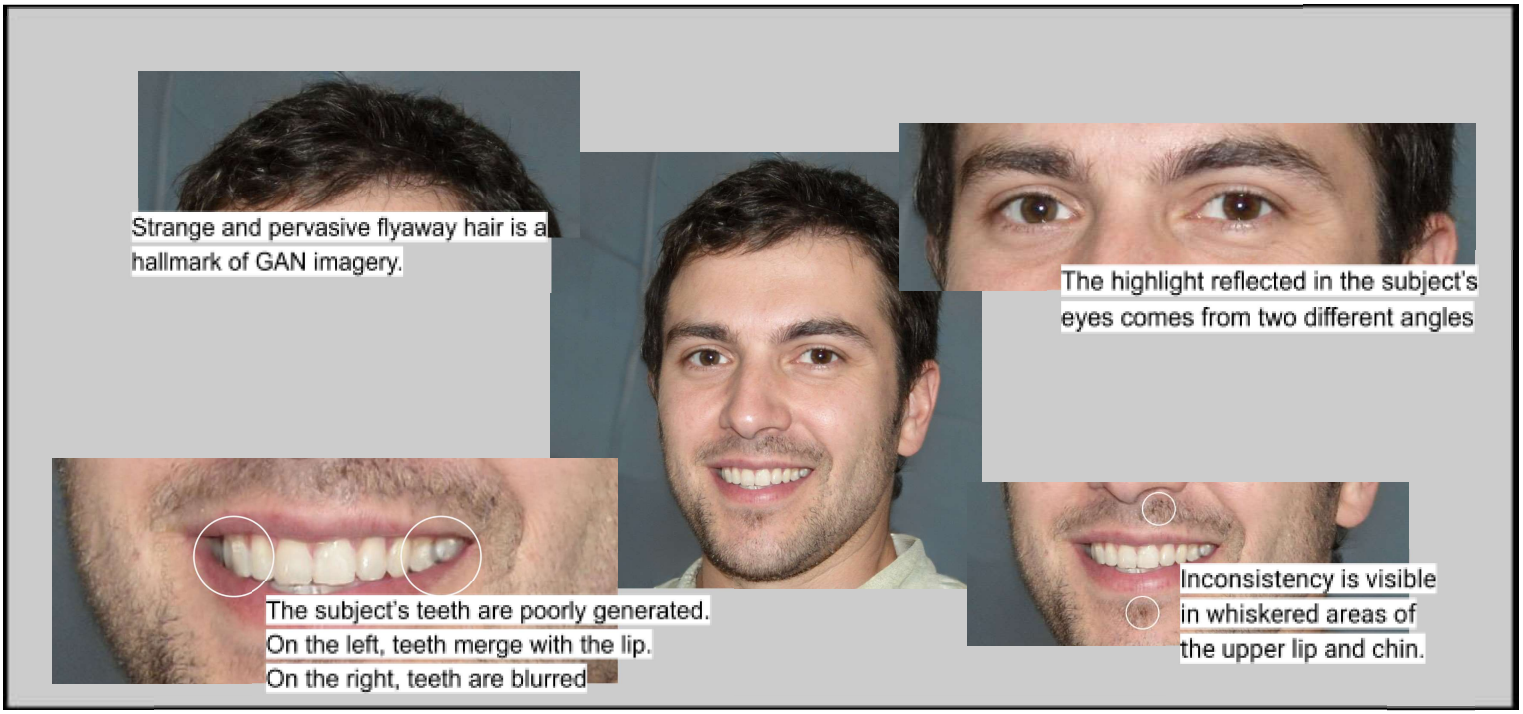
Figure 15

Discoloration in Generated DeepFake

Note. Example from FaceForensics [33] showing shading artifacts arising from illumination estimation and imprecise geometry of the nose. Taken from Matern, F., Riess, C., & Stamminger, M. (2019). Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations. *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 83-92.

Figure 16

Synthesis DeepFake with Visual Artifacts: Inconsistent Lighting, Unnatural Hair, and Unnatural Teeth



Note. Compiled from Hartman, T., & Satter, R. (2020, July 15). *These faces Are Not Real: How to Detect DeepFake Faces*. Reuters. https://graphics.reuters.com/CYBER-DEEPFAKE/ACTIVIST/nmovajgnxpa/

2.1 Technical Detection Approaches

Technical approaches for detecting DeepFakes target a variety of different weaknesses in DeepFake generation. The following section categorizes technical DeepFake detection models into blending artifact detection, environmental artifact detection, forensic artifact detection, behavioral artifact detection, physiological artifact detection, coherence based detection, anomaly detection, and generic classifiers. Specific details on detection models are described in Table 4.

Blending Artifact Detection:

Researchers have identified spatial blending artifacts on face-swap DeepFakes where the boundaries of facial images are semantically inconsistent when the image is replaced in the frame. This results in dissimilarity between neighboring pixels. A number of detection models have used local feature descriptors and frequency analysis to classify media as real or fake by comparing the similarity of pixels (Agarwal et al., 2017; Zhang, Zheng, & Thing., 2017; Akhtar & Dasgupta, 2019; Durall et al., 2019). Agarwal et al. note that while blending procedures in face-swap generation leave center regions well-blended, regions around eyes, nose and mouth tend to be vulnerable to artifacts (Agarwal et al., 2017). Liu & Lyu note the presence of residuals leftover from face-warping processes in lower resolution DeepFakes (Liu & Lyu, 2019).

Environmental Artifact Detection:

Shown in Figure 17, artifacts of face-swap DeepFakes include semantic inconsistencies between a face and its background. Researchers have used both patch and pair convolutional neural networks and encoder decoder networks to classify media based on discrepancies between foreground and background features (Li et al., 2020; Nirkin et al., 2020). Additionally, DeepFake content is prone to inconsistent lighting patterns. Straub specifically targets this inconsistency in his model, which makes both pixel-to-adjacent-pixel and regional lighting comparisons to differentiate authentic and DeepFake media (Straub, 2019).

Figure 17

Discrepency Between DeepFake Faces and Context Including Glasses, Hair, Ears and Neck

Note. Two example fake (swapped) faces from DFD. Left: The arm of the eyeglasses does not extend from face to context. Right: An apparent identity mismatch between face and context … these and similar discrepancies can be used as powerful signals for automatic detection of swapped faces. Taken from Nirkin, Y., Wolf, L., Keller, Y., & Hassner, T. (2020). DeepFake Detection Based on the Discrepancy Between the Face and its Context. *ArXiv:2008.12262 [Cs]*. http://arxiv.org/abs/2008.12262

Forensic Artifact Detection:

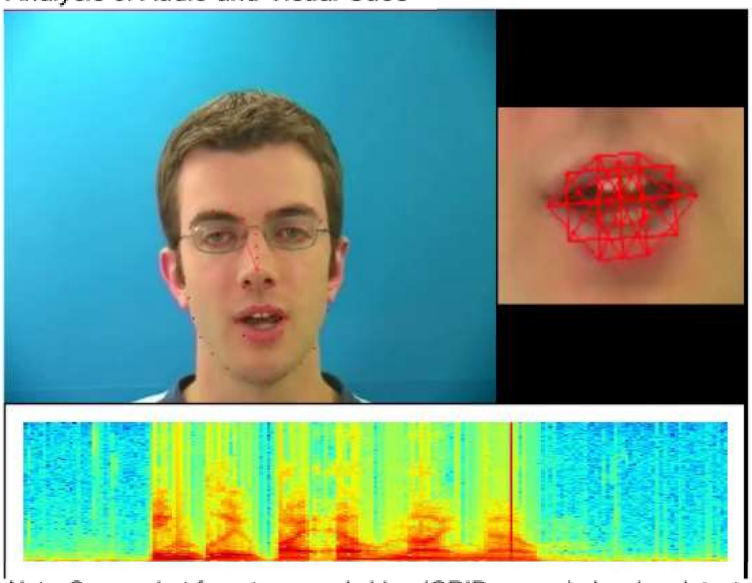
Forensic analysis has allowed researchers to identify manipulated media through subtle details in content. For example, Koopman et al. leverage patterns of sensor noise left by factory defects in digital cameras to differentiate real and DeepFake media (Koopman et al., 2018). Other forensic analysis has identified the fingerprints of pixel patterns generated by different generative adversarial networks (Marra et al., 2018).

Behavioral Artifact Detection:

Other methods of DeepFake detection target anomalies in the behavior of subjects. Agarwal et al. propose a model that learns the facial expression and speech patterns of world leaders such that DeepFakes can be identified by divergence in these learned patterns (Agarwal et al., 2019). Other authors leverage discrepancies in audio and visual cues (Mittal et al., 2020; Korshunov & Marcel, 2018; Korshunov et al., 2019).

Figure 18

Analysis of Audio and Visual Cues



Note. Screenshot from tampered video (GRID corpus) showing detected visual features and spectrogram. Taken from Korshunov, P., & Marcel, S. (2018). Speaker Inconsistency Detection in Tampered Video. *2018 26th European Signal Processing Conference (EUSIPCO)*, 2375–2379. https://doi.org/10.23919/EUSIPCO.2018.8553270

Physiological Artifact Detection:

Natural physiological patterns are also used to differentiate between forged and authentic content. Researchers have explored the use of pulse, heart rate and blinking as biological indicators of authenticity. Models are trained to classify media content based on these signals (Ciftci & Demir, 2020; Ciftci et al., 2020; Conotter et al., 2014; Li, Bao, et al., 2018)

Coherence Based Detection:

Due to a weakness of DeepFake generation methods at producing temporally coherent video footage, a number of detection approaches leverage classifiers that evaluate the temporal coherence of input media. By comparing video frames and identifying artifacts such as flicker and jitter, models are able to differentiate between authentic and fake content (Güera & Delp, 2018; Sabir et al., 2019; Amerini et al., 2019).

Anomaly Detection:

Other researchers have leveraged unsupervised deep learning architectures to recognize anomalies indicative of DeepFake content. These models are trained on normal data and detect deviations from authentic media patterns. Khalid and Woo compute an anomaly score of encoded and reconstructed images from a reconstruction network trained solely on real faces (Khalid & Woo, 2020). Other researchers have measured anomalies using facial recognition networks. This is accomplished through monitoring neural activation or analyzing how well an image fits training distributions (Want et al., 2020; Fernandes et al., 2020).

Generic Classifiers:

Unsupervised deep learning architectures are also deployed in generic classification models. One strength of deep learning based detection is better performance on compressed imagery (Marra et al., 2018). A number of authors propose models that use convolutional neural networks to classify input as real or fake (Afchar et al., 2018; Do Nhu et al., 2018; Tariq et al., 2018; Ding et al., 2019). Advances in the use of convolutional neural network classifications include Hsu et al's. use of Siamese convolutional neural networks to classify content (Hsu et al., 2020). Given that convolutional neural networks are blind to attacks that they are not trained on, Fernando et al. propose a Hierarchical Memory Network that utilizes neural memories to anticipate future semantic embeddings (Fernando et al., 2019). To produce a robust model that is less prone to false positives, Rana & Sung propose an ensemble learning technique that utilizes 7 distinct convolutional neural DeepFake detection networks (Rana & Sung, 2020). To exploit temporal weaknesses in DeepFake generation, de Lima et al. employ a 3D convolutional neural network to analyze multiple frames simultaneously (de Lima et al., 2020). However, it is noted

that generic classifiers are especially prone to adversarial machine learning attacks (Mirsky & Lee, 2022).

Table 4

Survey of DeepFake Technical Detection Models

Authors	**Detection Type**	**Description**
(Agarwal et al., 2017)	Blending Artifact Detection	* leverages differences in neighboring pixels * compares the difference between a center pixel and its neighbors * analysis through a Weighted Local Binary Pattern, which assigns weights inversely proportional to distance from a center pixel
(Mo et al., 2018)	Blending Artifact Detection	* leverages statistical analysis of pixels * pass input through a high pass filter to obtain residuals * use a convolutional neural network to classify images as real and fake
(Li, Bao, et al., 2020)	Blending Artifact Detection	* trains a convolutional neural network explicitly on blending boundaries in order to classify images as real and fake
(Li & Lyu, 2019)	Blending Artifact Detection	* analyzes images for residuals leftover from face-warping processes
(Li, Yu, et al., 2020)	Environmental Artifact Detection	* uses a patch and pair neural network to classify images based on differences in pixel distribution between face patches and background patches
(Nirkin et al., 2020)	Environmental Artifact Detection	* compares a face to its context. Context can include features such as hair, ears, and neck * processes faces and context separately in face encoder networks * decoder network is used to classify images as real or fake

(Straub et al., 2019)	Environmental Artifact Detection	* leverages both pixel-to-adjacent pixel comparisons and regional lighting comparisons to identify DeepFake media
(Yang et al., 2018)	Forensic Artifact Detection	* uses a support vector machine to classify images based on 3D head pose estimation
(Marra et al., 2018)	Forensic Artifact Detection	* identify DeepFake content based on fingerprints left by GAN noise residuals
(Yu et al., 2019)	Forensic Artifact Detection	* classify content based on unique fingerprints of GAN models
(Koopman et al., 2018)	Forensic Artifact Detection	* utilizes photo response non uniformity (PNRU) patterns * classifies based on normalized cross correlation scores of PNRU for video frames *PNRU refers to the noise pattern left by factory defects in digital cameras*
(Agarwal et al., 2019)	Behavioral Artifact Detection	* analyzes the expression behaviors of a specific identity through tracking facial movements, head movements, and muscle movements, which are encoded as action units * characterizes an individual's motion signature through Pearson correlation * trains a support vector machine on an identity's expression behavior to classify videos
(Mittal et al., 2019)	Behavioral Artifact Detection	* uses a Siamese network-based architecture to extract and analyze emotional cues of audio and visual content * classify based on similarity between audio and visual emotional cues

continued on next page

(Korshunov & Marcel, 2018)	Behavioral Artifact Detection	* detects audio-visual inconsistencies * compare performance of different feature processing methods * conclude that long short-term memory networks (LSTM) perform best *LSTM are recurrent neural networks that learn long term dependencies to better handle sequences of data*
(Korshunov et al., 2019)	Behavioral Artifact Detection	* use a two-class classifier to differentiate real and fake media based on inconsistency in audio and visual cues * utilize feature embeddings from a deep neural network trained on speech recognition for audio cues * utilize face and mouth landmarks for visual cues * use a long short-term memory network to learn temporal sequences of a video
(Ciftci & Demir, 2020)	Physiological Artifact Detection	* uses a convolutional neural network based classifier to label content as real or fake based off of analysis of biological signals in different facial regions
(Ciftci et al., 2020)	Physiological Artifact Detection	* use spatiotemporal patterns of biological signals to classify content as real or fake * fingerprint DeepFake source models by interpreting spatiotemporal biological signal patterns as a projection of residuals
(Connotter et al., 2014)	Physiological Artifact Detection	* differentiate computer generated and human faces based off of fluctuations in human faces due to changes in blood flow * these fluctuations are indicative of pulse

(Li, Bao, et al., 2018)	Physiological Artifact Detection	* utilizes a long short term recurrent CNN (LRCN) to analyze temporal states of eyes to analyze eye-blinking patterns * classify content based on presentation of natural eye-blinking *LRCNs are a combination of CNN and recursive neural networks. These networks are suited to handle temporal knowledge*
(Güera & Delp, 2018)	Coherence Based Detection	* train a recurrent neural network to recognize temporal artifacts, such as flicker and jitter, that are indicative of DeepFake content
(Sabir et al., 2019)	Coherence Based Detection	* leverage weakness in the temporal coherence of generated DeepFakes * utilize combinations of recurrent convolutional neural networks to identify temporal discrepancies
(Amerini et al., 2019)	Coherence Based Detection	* uses optical flow fields to identify dissimilarity between frames in a video * inter-frame dissimilarities are used as a feature in a CNN based classifier
(Khalid & Woo, 2020)	Anomaly Detection	* train a reconstruction variational autoencoder on only real faces * identify DeepFakes by passing content into the trained reconstruction network and computing an anomaly score * the anomaly score is based off of the mean square error of encoded and reconstructed images
(Wang et al., 2020)	Anomaly Detection	* monitor layer-by-layer neural activation of facial recognition models * uses neural coverage and activation to classify content as authentic or fake * robust against some adversarial perturbation attacks

(Fernandes et al., 2020)	Anomaly Detection	* proposes an attribution based confidence metric * classify images as real if confidence values are above 0.94
Afchar et al., 2018	Generic Classifier	* uses a deep learning method to classify images as real or fake * uses a low number of layers to focus on mesoscopic properties of images
(Do Nhu et al., 2018)	Generic Classifier	* use a deep convolutional neural network to detect DeepFakes * utilize a deep face recognition system for face feature extraction * network is fine tuned for real/fake image classification
(Ding et al., 2019)	Generic Classifier	* use deep transfer learning for face swap detection * include uncertainty measure with each prediction
(Hsu et al., 2020)	Generic Classifier	* train a fake feature network using pairwise deep learning to differentiate the features of real and fake images * add a classification layer to the fake feature network to label images as real/fake
(Fernando et al., 2019)	Generic Classifier	* use a convolutional neural network based approach for detecting face tampering * utilize a Hierarchical Memory Network architecture that stores information in neural memories and uses visual cues to predict future semantic embeddings * more robust to unseen manipulation techniques
(Rana & Sung, 2020)	Generic Classifier	* ensemble learning technique that uses 7 distinct convolutional neural networks trained for DeepFake detection * less prone to false positives

(de Lima et al., 2020)	Generic Classifier	* employ a 3D convolutional neural network to analyze frames simultaneously when classifying fake/authentic content
(Tariq et al., 2018)	Generic Classifier	* focus a neural network based classifier on image contents to detect forged faces

Weaknesses of Technical Detection Approaches:

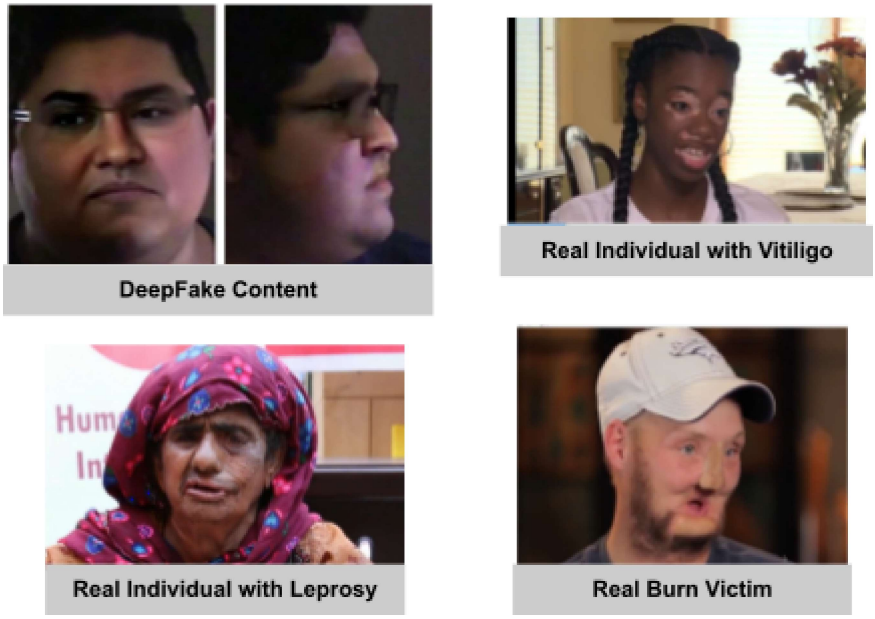
Masood et al. note that many existing detection mechanisms are best suited for face swaps. Lip-sync and expression manipulations leave more subtle artifacts that challenge existing detection architectures. These scholars note that research approaches have demonstrated greater reliability for image-based manipulation detection as compared to video-based decisions (Masood et al, 2021). Limits in DeepFake detection generalizability are related to the strong reliance of existing detection models on a finite set of research datasets (Pu et. al, 2021; Masood et al, 2021). The artifacts present in these training sets are not guaranteed to represent the artifacts present in deployed DeepFakes.

Performance of Technical Detection Across Different Communities:

Through an exploration of racial bias in detection models, Pu et al. conclude that the CapsuleForensics detection model has the highest accuracy on classifying the authenticity of videos for input featuring Black faces. The F1 score of classification for this group is 74%. The performance on Caucasian faces is comparable with an F1 score of 72%. However, the performance on Asian faces drops to a mere 48% (Pu et al., 2021). Other ethno-racial categories were not investigated. During the 2020 CVPR Media Forensics Workshop, Prabhu et al. commented on populations whose videos were likely to experience a high degree of false positive classification in DeepFake detection models. Detection based on blending artifacts is prone to misclassifying the faces of individuals who have conditions such as leprosy or vitiligo. Detection based on blending artifacts is likely to misclassify the faces of burn victims, individuals with facial tattoos and smooth baby faces (Prabhu, 2020).

Figure 19

Communities Vulnerable to DeepFake Misclassification



Note. Compiled from Prabhu, A., Materzyńska, J., Dokania, P. K., Torr, P. H. S., & Lim, S.-N. (2020, June 15). *Is Your Face Fake? Social Impacts of Algorithmic "Fake" Determination* [Conference]. DFDC Risk-a-thon & CVPR media forensics workshop 2020, Seattle Washington. https://drimpossible.github.io/documents/dfdc_slides.pdf

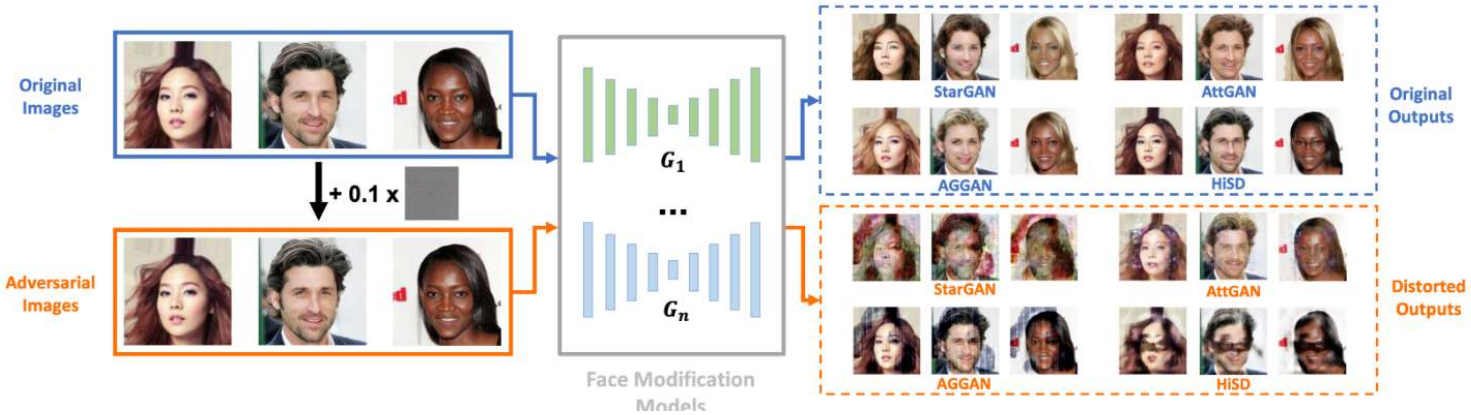
2.2 Adversarial Image Perturbation

One preventative measure against DeepFake generation is adversarial image manipulations that target weaknesses in DeepFake generation models. A number of different models have been proposed to perturb images. Yeh et al propose a method that applies adversarial loss to images in such a way that manipulating these images is rendered more difficult. This adversarial attack specifically targets image translation models such as CycleGAN, pix2pix and pix2pixHD (Yeh et al., 2020). Segalis and Galili propose a model that targets face-swapping autoencoders. This OGAN model iteratively trains an adversarial image generator against a face-swapping model to create a model of training resistant adversarial image perturbations. The model is more robust toward DeepFake generation models trained on datasets that include adversarially manipulated input images (Segalis & Galili, 2020). Dong & Xie explore 3 different adversarial attacks on autoencoders. One universal image perturbation model is image agnostic. The other two models provide precise, image-specific distortions (Dong & Xie., 2021). Huang et al. propose a robust Cross-Model Universal Watermark that protects a variety of facial images from multiple DeepFake models. This attack iteratively trains attacks

against multiple DeepFake models. Then, the authors propose a two-level processing step to reduce conflicts between resulting watermarks (Huang et al., 2021).

Figure 20

Adversarial Perturbations as Protection Against GAN-Based Manipulations



Note. Illustration of the CMUA-Watermark. Once the CMUA-watermark has been generated, it can be directly added to any facial image to generate a protected image that is visually identical to the original image but can distort outputs of deepfake models. Taken from Huang, H., Wang, Y., Chen, Z., Zhang, Y., Li, Y., Tang, Z., Chu, W., Chen, J., Lin, W., & Ma, K.-K. (2021). CMUA-Watermark: A Cross-Model Universal Adversarial Watermark for Combating Deepfakes. *ArXiv:2105.10872 [Cs]*. http://arxiv.org/abs/2105.10872

2.3 Distributed Ledger Technologies

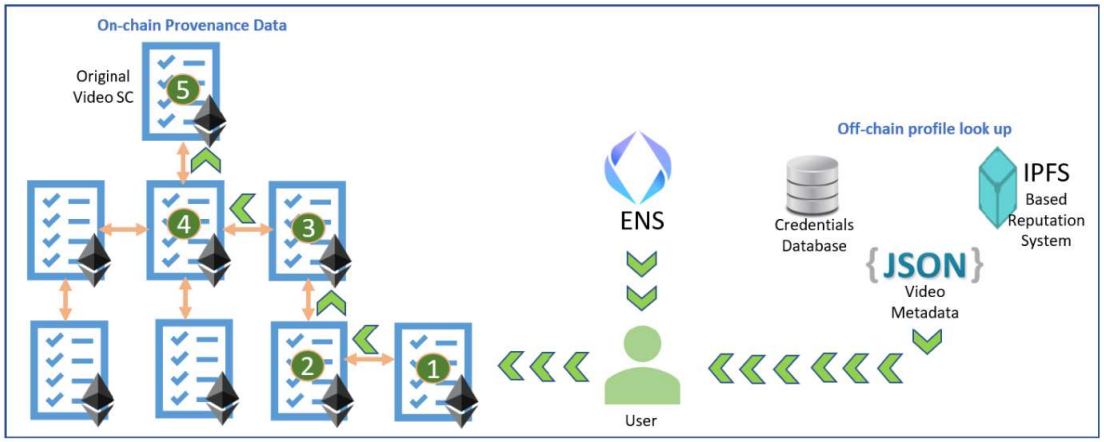
Provenance Based Approaches:

In 2019, Hasan and Salah proposed a framework of digital provenance and history tracking to combat the threat of DeepFake attacks. This framework uses blockchain technologies to provide credible and secure proof of authentication through traceability to a trusted data source. These authors leverage features of the InterPlanetary File System (IPFS) decentralized storage, a decentralized reputation system, and Ethereum Name service. The authors note that one challenge of provenance based solutions is establishing trust in a signing authority (Hasan & Salah, 2019). The code for Hasan & Salah's framework is publicly available on GitHub (smartcontract694, 2018/2022). England et al. propose an alternative framework to track media provenance.This authentication is characterized by a system of verified manifests. When media is uploaded by a content provider, a publisher-signed manifest is created. This manifest is registered and signed by a permissioned ledger authority via the Confidential Consortium Framework (CCF). Manifests are stored in a database that allows for fast lookup via web browser (England et al., 2021). To inform the design of provenance indicators, Sherman et al.

conduct user interviews. These interviews reveal that media provenance is a key heuristic leveraged by users to identify misinformation (Sherman et al., 2021).

Figure 20

Blockchain Based History Tracking for Media Provenance



Note. Tracing video source origin using [Hasan & Shah's] proposed solution.Taken from Hasan, H. R., & Salah, K. (2019). Combating Deepfake Videos Using Blockchain and Smart Contracts. *IEEE Access*, 7, 41596–41606. https://doi.org/10.1109/ACCESS.2019.2905689

Content Moderation Approaches:

Other applications of distributed ledger technologies to combat deceptive media hone in on content moderation. Frameworks apply blockchain technologies to decentralized content moderation, trustworthiness checkers, incentivized fact checking, decentralized social media platforms, and reputation systems. Trustworthiness checkers are social networks that allow any node (ie. user) to verify that content is truthful. Fact-checking incentivized applications utilize reputation metrics to incentivize the reliability of fact-checking behavior through monetary rewards for reliable fact-checkers. Reputation systems produce credibility scores for the publishers of content (Fraga-Lamas & Fernández-Caramés, 2020). For distributed content moderation to succeed, users must be digitally literate.

Digital Literacy

Other authors have argued the need for greater digital literacy among internet users (Westerlund, 2019). By preparing users to anticipate the presence of deceptive media, visual and auditory content can be more critically consumed. While awareness of distinguishing visual

artifacts allows users to identify some unsophisticated DeepFakes, digital literacy more significantly employs heuristics to evaluate the credibility of information sources and content.

3. Current State of DeepFake Detection and Generation Arms Race

3.1 Kaggle DeepFake Detection Challenge

Dataset Development:

In 2020, Facebook AI, AWS, Microsoft and the Partnership on AI Steering Committee partnered with Kaggle to host an open competition of DeepFake face swap detection models (Kaggle, 2020). For this competition, researchers at Facebook AI developed a novel dataset containing more than 100,000 videos for use as a blackbox test set for challenge submissions. This dataset was developed with footage from 3,426 consenting, paid actors and eight different facial manipulation algorithms. The authors recognized that existing DeepFake datasets had overrepresentation of actors in non-natural settings, such as news and briefing rooms, which lead to underrepresentation of natural illumination conditions in research datasets. To fill this deficit, the authors of the dataset staged videos in a variety of different natural lighting conditions.

Research Commentary On Generation Weaknesses:

The developers of this dataset note that DeepFake autoencoders, convolutional autoencoders with one shared encoder and two identity specific decoders, provided flexible DeepFake generation under a variety of lighting conditions. However, shown in Figure 21, this architecture had weaknesses around extreme poses and glasses (Dolhansky et al., 2020).

Figure 21

DeepFake Generation Using DeepFake Autoencoder Architecture

Note. A selection of results of varying quality. Quality increases from left to right. Taken from Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). The DeepFake Detection Challenge (DFDC) Dataset. *ArXiv:2006.07397 [Cs]*. http://arxiv.org/abs/2006.07397

Additionally, frame-based morphable mask models tended to work well on single-frame images, but produced discontinuities in the face and occasionally failed to fit the mask to a face as shown in Figure 22 (Dolhansky et al., 2020).

Figure 22

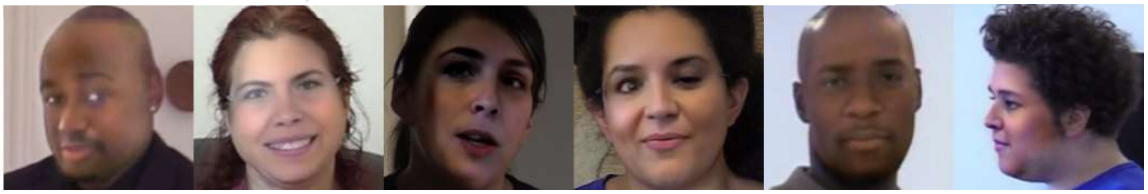
DeepFake Generation Using Frame Based Morphable Mask Models

Note. A selection of results of varying quality. Quality increases from left to right. Taken from Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). The DeepFake Detection Challenge (DFDC) Dataset. *ArXiv:2006.07397 [Cs].* http://arxiv.org/abs/2006.07397

Shown in Figure 23, the FSGAN model functioned well in good lighting conditions and translated extreme poses well. However, it experienced poor performance in dark lighting conditions (Dolhansky et al., 2020).

Figure 23

DeepFake Generation Using FSGAN



Note. A selection of results of varying quality. Quality increases from left to right. Taken from Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). The DeepFake Detection Challenge (DFDC) Dataset. *ArXiv:2006.07397 [Cs].* http://arxiv.org/abs/2006.07397

The GAN based neural talking head model had fairly consistent performance, but performed poorly in poor lighting conditions. Additionally, seen in Figure 24, the model produced visually similar eyes on all DeepFake generations (Dolhansky et al., 2020).

Figure 25

DeepFake Generation Using A GAN Based Neural Talking Head Model

Note. A selection of results of varying quality. Quality increases from left to right. Taken from Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). The DeepFake Detection Challenge (DFDC) Dataset. *ArXiv:2006.07397 [Cs].* http://arxiv.org/abs/2006.07397

Finally, researchers observed poor performance of the StyleGAN model with weaknesses of semantically invalid eye poses, such as eyes looking in different directions, and mismatched illumination (Dolhansky et al., 2020). Generations using StyleGAN are shown in Figure 26.

Figure 26

DeepFake Generation Using StyleGAN

Note. A selection of results of varying quality. Quality increases from left to right. Taken from Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). The DeepFake Detection Challenge (DFDC) Dataset. *ArXiv:2006.07397 [Cs].* http://arxiv.org/abs/2006.07397

Challenge Results:

The submission to the 2020 Kaggle DeepFake Detection Challenge revealed that current face swap generation technologies are outpacing technical detection methods. Of the 21,114 submissions, the top performing model only achieved an average precision of 65% against the black-boxed dataset. This model ranked fourth in precision on the publicly available dataset. The best performance on the public test set reached a mere 83% average precision (Facebook AI, 2020). Performance on the private test set was poor across the board. 60% of submissions had log loss lower than or equivalent to randomly guessing on a balanced test set. Good performance on the public test set was correlated with good performance on the private test set.The top performing solution used a multi-task cascaded convolutional neural network for facial detection and alignment and an EfficientNet network for feature encoding. Many other top-performing solutions also used combinations of convolutional neural network architectures including EfficientNet networks and Xception architectures (Dolhansky et al., 2020; Tan & Le, 2020; Chollet, 2017).

DeepFake Videos In The Wild: Analysis and Detection

Summary of Results:

In 2021, a collaboration of researchers at Virginia Tech, the University of Virginia, the University of Michigan, Facebook and LUMS Pakistan produced an analysis of state of the art DeepFake detection model performance on a DeepFake video test set created from a collection of non-pornographic DeepFakes. These videos were found on online platforms such as Youtube, Billibilli and Reddit using targeted search queries and DeepFake specific subforums. The 7

tested detection models performed poorly on the DeepFakes In the Wild dataset. The best performing model, CapsuleForensics, which employs both a VGGFace network and a Capsule network, had an F1 score below 77%. The worst performing model, Multitask, built using a multi-output autoencoder, only achieved an F1 score of 66%. All models had precision below 69%, which indicated the presence of false positives. The authors conclude that detection does not generalize well to DeepFakes found in the wild. Contrary to their hypothesis, they observed comparable performance between supervised and unsupervised detection models (Pu et al., 2021).

Weakness in Research Dataset Representation:

In their discussion, Pu et al. make a number of observations on the current state of the arms race between DeepFake generation and detection. The authors attempt to identify the generation method of the videos in their dataset and find that 94.2% of videos found on Youtube were generated using DeepFaceLab software. They note that no existing research datasets have representation of videos produced using DeepFaceLab software, despite its high prevalence in the wild. This lends to a greater claim that the datasets used by the research community are not necessarily representative of the DeepFakes produced in the wild. To allow for more representative and specific DeepFake detection, the authors propose a Deep Neural Network to fingerprint the model used to create a DeepFake. This proposed network leverages a fingerprinting model that is trained to fingerprint a GAN model from a GAN-generated image (Pu et al, 2021).

Weakness In Detection Assumptions:

The authors also identify a number of assumptions made by DeepFake detection models that do not hold up to DeepFakes in the wild (Pu et. al, 2021). For example, many detection models assume that every frame of a video has a fake face. In the wild, this was not found to be true. Additionally, models are designed towards DeepFakes with one face in each frame. DeepFakes in the wild were found to contain multiple faces in a frame. Furthermore, DeepFakes in the wild tend to have a longer duration than DeepFakes found in research datasets. The authors note that this leads to a weakness where DeepFakes videos with a large number of clean frames are likely to be falsely classified as non-DeepFake content, since the classification is often determined via an average of frame scores. The authors argue for a classification method first proposed by Li & Lyu by which a top percentile of frame scores are used to compute a classification (Li & Lyu, 2019; Pu et al., 2021).

Weaknesses to Adversarial Attacks:

Curious as to which features are identified as relevant by detection schemes, the authors utilize IntGrad, a DNN based feature-attribution explanation methodology to analyze detection models (Sundararajan et al., 2017; Pu et al, 2021). They find that detection models are more likely to identify an image with more background features as real, which allows adversaries to pass in DeepFakes with background noise to spoof detection models. Pu et al. argue that identifying facial boundaries and confining analysis to relevant regions is critical for accurate DeepFake detection (Pu et al., 2017). Other researchers have also investigated the weaknesses of detection methods to adversarial attacks. In 2021, Fan et. al proposed a Poisson noise DeepFool model that iteratively develops adversarial examples. In experiments, this model weakened DeepFake detection accuracy from 0.9997 to 0.0731 (Fan et. al, 2021).

4. Games of Cat and Mouse

While the detection of DeepFakes poses unique challenges, it is not the first situation where the technology of malicious adversaries has been caught in a game of cat and mouse with complementary defensive technologies.

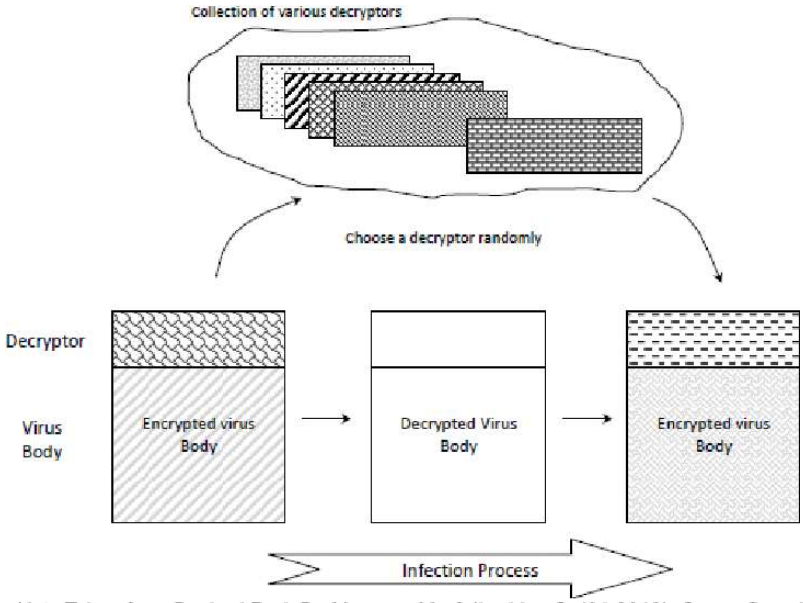
4.1 A History of Malware Generation and Anti-Malware Detection

The competition between computer malware and anti-malware technologies can be traced back to 1987 with the insertion of a Trojan horse into Ross Greenberg's Flushot IV antivirus program. In response, Greenberg developed Flushot Plus (Marshall, 1988). Preliminary approaches to writing anti-virus software utilized simple signature detection methods. Signature detection code identifies the presence of malicious malware by matching bytes of executable code to known virus signatures. In response to the deployment of signature-based antivirus, virus writers began to encrypt their viruses such that the code body no longer matched a given virus signature. The first encrypted virus was the DOS virus CASCADE developed in 1988 (Rad et al., 2011).

When anti-virus began detecting signatures for encrypted viruses, virus writers obfuscated their viruses through mutation. Oligiomorphic viruses utilize a set of varying decryptor loops so that not all infections by a particular virus are identical. This adds additional overhead to the process of signature scanning. With oligiomorphic viruses, it is necessary to identify multiple signatures for a singular virus (Rad et al., 2011). Oligiomorphic viruses prompted the development of more efficient virus scanners through techniques such as hashing, top and tail scanning and generic signatures with flexibility from mismatches and wildcards.

Figure 27

Structure of Oligiomorphic Viruses

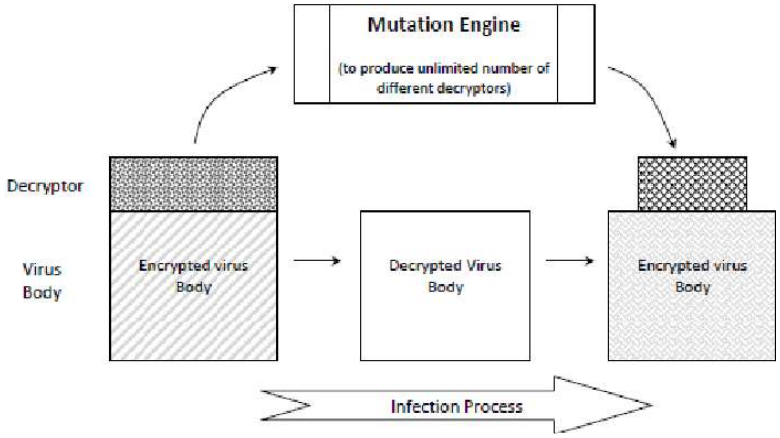


*Note.*Taken from Bashari Rad, B., Masrom, M., & Ibrahim, S. (01 2012). Camouflage In Malware: From Encryption To Metamorphism. *International Journal of Computer Science And Network Security (IJCSNS)*, *12*, 74–83.

To evade efficient virus scanners, virus writers developed polymorphic viruses which mutate the decryptor with each new infection. One virus scanning advancement, X-RAY scanning targets weaknesses in virus encryption to allow for plain text scanning. Metamorphic viruses were developed to evade advancements in antivirus by mutating not only decryption code, but instead mutating the virus body with each new infection.

Figure 28

Structure of Polymorphic Viruses



*Note.*Taken from Bashari Rad, B., Masrom, M., & Ibrahim, S. (01 2012). Camouflage In Malware: From Encryption To Metamorphism. *International Journal of Computer Science And Network Security (IJCSNS)*, *12*, 74–83.

Ultimately, rather than improve virus signature scanners, antivirus engines moved towards virtualization that protects computer hardware. Code emulation runs executable code on virtual hardware and waits for a polymorphic virus to decrypt itself before scanning. However, in response to this strategy, advanced viruses began to detect virtual environments and will stop execution if emulation is suspected (Rad et al., 2011).

4.2 Computational Advantage of Malware Generation

The coevolution of malware and anti-malware technologies is not a balanced arms race. Detecting malware is more challenging and expensive than developing virus code (Menéndez et al., 2021). Malware writers have more limited scope when infecting a program; the goal to embed malicious code in a target program can be accomplished through a number of different avenues and must merely exploit finite vulnerabilities. The behavior of antivirus and common computer programs is defined. On the other hand, Fred Cohen proved in 1987 that it is theoretically impossible to write an algorithm to perfectly detect all computer viruses (Cohen, 1987).

To protect against all attacks, antivirus would need to not only protect against existing cyber threats, but also provide protection against unknown and novel threats. In addition, perfect anti-virus would require proof that every executable program on a computer does not contain malicious code. The process of scanning every executable program would degrade system performance. Antivirus writers recognize the infeasibility of perfect virus detection and instead balance a number of efficiency and accuracy trade-offs to produce sufficiently desirable performance. Rad et al. note that users will not purchase antivirus engines that produce too many false positives (Rad et al., 2011). When antivirus systems quarantine benign programs, users face inconvenience. Additionally, antivirus systems are incentivized to limit the resources and time for which they run. From a user perspective, antivirus systems that consume too many resources and slow down other programs are not desirable. For this reason, antivirus systems use a number of heuristics to merely scan and analyze portions of executable files that are likely to contain virus code (Rad et al., 2011).

4.3 Computational Advantage of DeepFake Generation

Technical Detection:

DeepFake generation and detection both rely on complex machine learning models that require access to graphical processing units and a significant training overhead. However, the scope of the media to which each is applied varies greatly. DeepFake generation targets finite use cases and need only train towards the production of finite media for target identities. In contrast, perfect DeepFake detection would require flexible identification of any falsified media, which is

a much broader domain. DeepFake generation can involve tedious, iterative cycles to improve fidelity. However, DeepFake detection must balance constraints of computational resources to remain scalable to practical use. On average, 500 hours of video footage per minute are uploaded to Youtube (Bernaciak & Ross, 2022). Digital platforms, such as Youtube and Billibilli, have the power to enforce constraints on uploads. However, given that top performing DeepFake detection algorithms require access to graphical processing units and sufficient memory, DeepFake detection cannot be hosted on local machines that lack these resources (Hao, 2020; Seferbekov, 2020/2022). Technology platforms have sufficient resources to implement DeepFake detection, but will likely balance DeepFake detection into a network of computational and performance costs. Given the current framework of misinformation policies on technology platforms, DeepFake detection is likely to first see use as a data point in more complicated user-initiated content moderation procedures.

Provenance Based Solutions:

Recent research proposals have acknowledged that there are logistic feasibility challenges to widespread adoption of mitigation solutions. Provenance based solutions are gaining traction in the research and legislative communities (Lima, 2021). Proponents of these strategies recognize feasibility constraints and frictions. Dhal et. al discuss the network scalability design considerations of their blockchain and keyed watermark based framework for provenance on social media (Dhall et al., 2021). England et al. conduct experiments to demonstrate that their proposed Authentication of Media via Provenance (AMP) ledger system scales well for HTTP Adaptive Streaming. The observed latency threshold was low enough to not interfere with user viewing experience (England et al., 2021). Other scholars recognize that the success of provenance solutions does not necessarily rely on ubiquitous adoption of ledger technologies, but rather a system where verified content can be traced to trusted news and media authorities (Aythora et al., 2020). For provenance approaches to succeed at combating DeepFake misinformation, there is a necessary level of digital literacy and skepticism that users must exhibit to question information. From survey data, Sherman et al. conclude that users view provenance as an important heuristic for determining the reliability of media (Sherman et al., 2021). This gives some weight to the applicability of provenance based approaches. However, it is important to note that changes to protocols are a historically slow process, due to contention over benefits, trade-offs and backwards compatibility (Handley, 2006). Any adoption of provenance based approaches is unlikely to start with broad adoption.. Rather, if practical adoption of provenance is seen, it is likely to start with verified organizations such as news outlets.

4.4 Do You Know Your Enemy - Competitive Advantage of Knowledge on the Adversary

Zero Day Vulnerabilities:

In the context of malware, there is a concept known as a zero day vulnerability. This is a vulnerability that has not been discovered by benevolent actors and instead is at risk of exploitation by malicious adversaries. The associated concept of zero day exploits refers to attacks that exploit these overlooked vulnerabilities. Developers have zero days to patch the vulnerable software before it is exploited (FireEye, n.d.).

Advantage of Malware Writers:

Malicious actors have access to a number of forensics and information gathering tools to identify exploitable weaknesses. Adversaries can use network scanners, network traffic analysis, password cracking tools, vulnerability scanners, fuzzing tools, reverse engineering tools and other information collection tactics to develop exploitations. Any access to software allows for information gathering. Tools such as Fuzzdb contain prebuilt attack payloads that can be leveraged against unsecure systems (Fuzzdb-Project/Fuzzdb, 2015/2022). Fuzzing is a technique used by both software security professionals and malicious adversaries. In this automated process, variations of input are passed into a system with the intent of discovering exploitable vulnerabilities (Li, Zhao, et al., 2018). For example, fuzzing exploits can embed shell code into target programs, pass arguments to system calls, carry out SQL injection attacks, or reveal internal behavior of systems (MITRE, 2021). Malware writers have advantage in their ability to gather information on the weaknesses of the defenses used by their target.

Advantage of DeepFake Generators:

Many DeepFake detection methods have been made publicly available through GitHub repositories or published research papers. In the development of DeepFakes to circumvent existing detection methods, DeepFake developers have access to detection models and are able to test for vulnerabilities in existing detection architectures. DeepFake developers are able to train towards DeepFakes that fit a domain that is undetectable by existing detection models, but convincing to the human eye. There is no limit on the amount of input that DeepFake developers can pass into open source detection models. However, once a DeepFake exploits vulnerabilities to fail detection, those interested in detecting DeepFakes have zero days to discover that the DeepFake detection has failed before there is potential for negative implications.

Overfitting and Novel Threats:

Recent research has shown that deep learning based DeepFake detection methods are overfitted toward research community datasets and perform reasonably poorly on novel DeepFakes (Pu et al., 2021; Dolhansky et al., 2020). Just as malware writers learn to exploit vulnerabilities of computer anti-malware scanners, DeepFake generation methods develop techniques to better evade detection through learning weaknesses of existing detection techniques. In this way, the cat and mouse game between DeepFake generation and DeepFake detection is driven by generation techniques. This dynamic leaves DeepFake generation one step ahead of DeepFake detection. While general computer vision advances are able to aid the fine-grained visual classification techniques of DeepFake detection, the edge that DeepFake generation has over DeepFake detection is exacerbated by the fact that many of the top performing models of DeepFake detection rely on large sets of training data. Pu et al. show that the training sets that are popular in the research community are not representative of the DeepFakes found in the wild (Pu et al., 2021). Researchers must balance the representation of different DeepFakes in their training sets and stay up to date on recent developments as they attempt to catch increasingly more sophisticated DeepFake generations. Given the time and training data necessary to create quality DeepFakes, the procurement of state of the art DeepFake datasets is a limiting factor in the development of better detection algorithms. It is likely that DeepFake generation will continue to have an edge over DeepFake detection. Use of DeepFake detection methods may be a better heuristic tool to evaluate the authenticity of media than a catch-all tool to prevent DeepFake generation.

4.5 Beyond Code - Social Engineering Exploits

Social Engineering and Malware:

Malicious adversaries do not merely exploit vulnerabilities in technical software. Exploitations of user psychology are also used in the context of phishing, baiting and scareware. Baiting seeks to exploit user interest or curiosity. Examples of digital baiting attacks include malware masked as a desirable software or media download. A physical baiting attack can take the form of a USB drive left in a parking lot, sparking user interest (Paganini, 2020). Scareware exploits user anxieties. It can take the form of popup banners on a web browser that indicate the presence of computer viruses, prompting users to download malware that is disguised as antivirus (Stouffer, 2021).

Social Engineering and DeepFakes:

The attack surface of DeepFakes extends beyond a technical detection problem. DeepFakes can be used to introduce a level of psychological doubt that leaves viewers vulnerable to other social engineering tactics. Additionally, researchers note that users display a predisposition to more readily trust faces generated via generative adversarial network than real faces (Nightingale & Farid, 2022). In this way, the threat of DeepFakes cannot easily be solved through binary classification alone.

5. Semantic Context and Model Training Decisions

Not all DeepFake attack surfaces are created equally or warrant the same treatment. A 2019 web crawl by DeepTrace found a high prominence of non-consensual DeepFake pornography (Ajder et al., 2019). While the believability of DeepFake pornography adds danger to threats of blackmail and manipulation, the danger of the attack is less based on questions of indeterminate authenticity. Targets of DeepFake pornography videos can testify to the content's fake nature. However, this does not protect victims from violations of privacy, consent, defamation or legal repercussions. Additionally, there is no guarantee that a victim of such an attack will be believed. Supporting victims of non-consensual DeepFake pornography requires a deeper understanding of victim and viewer experience. An Instagram based phishing scam in India targeted victims by sending DeepFake pornography videos to the friends and family of the victim if the scammer did not recieve payment (Joshi, 2021). Combating the threat of these types of schemes involves greater public awareness to the threat of DeepFake pornography attacks. With greater awareness, the pornographic media can more easily be dismissed as fake. Additionally, in cases where victims find it beneficial to use detection technology to substantiate their claim to the forged nature of a pornographic video, detection algorithms should be biased towards falsely identifying real videos as fake rather than optimized for accuracy. In this context, false negatives produce greater harm to the subject of this media than false positives. Given that DeepFake detection algorithms tend to overfit to the data they are trained on, it would be reasonable for future research to specifically target model training towards the context of a face-swap video. For example, lighting conditions and speech patterns will be different in interview-based DeepFakes than pornographic DeepFakes. Model training that is specific to different attack surfaces allows for different trade-off decisions on false positive and false negative rates based on the situational risk of each outcome.

Conclusion

A variety of distinct deep learning architectures have been applied to the creation of forged media known as DeepFakes. These forgeries swap faces of source and target subjects,

manipulate features, drive the actions of a subject, manipulate audio, and synthesize new identities. Applications of generative adversarial networks, in tandem with variational autoencoders and 3D morphable mask models, are increasingly prevalent in this domain. Though generation has seen many improvements in recent years, DeepFake generation models still face weakness around generalizability, occlusions, identity leakage and temporal coherence.

DeepFake countermeasures have included a variety of technical detection models that leverage different weaknesses in DeepFakes. Supervised artifact based detection leverages both visual and temporal features. Visual features include blending artifacts, such as pixel dissimilarity, environmental artifacts, such as foreground and background incoherence, and forensic artifacts, such as pixel patterns from GAN procedures. DeepFakes are also identified through temporal features such as behavioral anomalies and lack of physiological indicators. DeepFake detection has deployed unsupervised machine learning models in generic classifiers and anomaly detection. Another technical countermeasure to DeepFakes is adversarial perturbations to images that impede DeepFake generation. A countermeasure that has recently gained a lot of traction is blockchain based provenance systems that allow users to trace media to its original source.

Currently, DeepFake generation technology outperforms technical DeepFake detection models. Weakness in DeepFake detection includes overfitting towards research community datasets, model assumptions that do not match real world DeepFakes, and vulnerability to adversarial attacks.

Given the precedent of malware generation and detection, it is likely that DeepFake generation will continue to outpace technical DeepFake detection. Parallel to malware detection, DeepFake detection has a much broader goal and is constrained by performance concerns within a larger system. DeepFake detection must be flexible to detect any forgery. DeepFake generation must only produce one forgery that fools detection. Just as malware writers are able to use fuzzing to identify vulnerabilities in antimalware systems, malicious actors hoping to improve DeepFake generation can test and fine-tune their generations towards open source DeepFake detection methods. Just as virus writers leverage social engineering to advance their attacks, malicious actors using DeepFakes can create and exploit user doubt.

Rather than attempt to win an unbalanced game, future research into DeepFake countermeasures should center user experience of DeepFakes to identify appropriate digital literacy and technical mitigation steps. Additionally, research can focus model training towards specific semantic contexts to best balance benefits and drawbacks of detection models.

Works Cited

AB-730 Elections: deceptive audio or visual media, Ca. (2019). https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB730

Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: A Compact Facial Video Forgery Detection Network. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, 1–7. https://doi.org/10.1109/WIFS.2018.8630761

Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., & Li, H. (2019). Protecting World Leaders Against Deep Fakes. *CVPR Workshops*.

Ajder, H., Patrini, G., Cavalli, F., & Cullen, L. (2019). The State of DeepFakes: Landscape, Threats, and Impact. *DeepTrace*, 27.

Akhtar, Z., & Dasgupta, D. (2019). A Comparative Evaluation of Local Feature Descriptors for DeepFakes Detection. *2019 IEEE International Symposium on Technologies for Homeland Security (HST)*, 1–5. https://doi.org/10.1109/HST47167.2019.9033005

Allyn, B. (2022, March 16). A Deepfake Video Showing Volodymyr Zelenskyy Surrendering Worries Experts. *NPR*. https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia

Amerini, I., Galteri, L., Caldelli, R., & Bimbo, A. (2019). Deepfake Video Detection through Optical Flow Based CNN. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. https://doi.org/10.1109/ICCVW.2019.00152

Aubé, T. (2017, February 13). AI, DeepFakes, and the End of Truth. *Medium*. https://medium.com/swlh/ai-and-the-end-of-truth-9a42675de18

Aythora, J., Burke, R., Chamayou, A., Clebsch, S., Costa, M., Earnshaw, N., Ellis, L., England, P., Fournet, C., Gaylor, M., Halford, C., Horvitz, E., Jenks, A., Kane, K., Lavallee, M., Lowenstein, S., MacCormack, B., Malvar, H., O'Brien, S., … Zaman, A. (2020). *MULTI-STAKEHOLDER MEDIA PROVENANCE MANAGEMENT TO COUNTER SYNTHETIC MEDIA RISKS IN NEWS PUBLISHING*. 11.

Ayyub, R. (2018, November 21). I Was The Victim Of A Deepfake Porn Plot Intended To Silence Me. *Huffington Post*. https://www.huffingtonpost.co.uk/entry/deepfake-porn_uk_5bf2c126e4b0f32bd58ba316

Bao, J., Chen, D., Wen, F., Li, H., & Hua, G. (2017). *CVAE-GAN: Fine-Grained Image Generation Through Asymmetric Training*. 2745–2754. https://openaccess.thecvf.com/content_iccv_2017/html/Bao_CVAE-GAN_Fine-Grained_Image_ICCV_2017_paper.html

Bickert, M. (2020, January 7). Enforcing Against Manipulated Media. *Meta*. https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '99*, 187–194. https://doi.org/10.1145/311535.311556

Bond, S. (2022, March 27). *The latest marketing tactic on LinkedIn: AI-generated faces: NPR*. https://www.npr.org/2022/03/27/1088140809/fake-linkedin-profiles

Bonner, A. (2019, June 1). *The Complete Beginner's Guide to Deep Learning: Artificial Neural Networks*. Medium.

https://towardsdatascience.com/simply-deep-learning-an-effortless-introduction-45591a1c4abb

Bregler, C., Covell, M., & Slaney, M. (1997). Video Rewrite: Driving visual speech with audio. *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '97*, 353–360. https://doi.org/10.1145/258734.258880

Brewster, T. (2021, October 14). Fraudsters Cloned Company Director's Voice In $35 Million Bank Heist, Police Find. *Forbes*. https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/

Brock, A., Donahue, J., & Simonyan, K. (2019). Large Scale GAN Training for High Fidelity Natural Image Synthesis. *ArXiv:1809.11096 [Cs, Stat]*. http://arxiv.org/abs/1809.11096

Chan, C., Ginosar, S., Zhou, T., & Efros, A. A. (2019). *Everybody Dance Now*. 5933–5942. https://openaccess.thecvf.com/content_ICCV_2019/html/Chan_Everybody_Dance_Now_ICCV_2019_paper.html

Chesney, B., & Citron, D. (2019). Deep fakes: looming challenge for privacy, democracy, and national security. *California Law Review, 107(6),* 1753-1820.

Choi, Y., Uh, Y., Yoo, J., & Ha, J.-W. (2020). StarGAN v2: Diverse Image Synthesis for Multiple Domains. *ArXiv:1912.01865 [Cs]*. http://arxiv.org/abs/1912.01865

Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. *ArXiv:1610.02357 [Cs]*. http://arxiv.org/abs/1610.02357

Ciancaglin, V., Gibson, C., Sancho, D., McCarthy, O., Eira, M., Amann, P., Klayn, A., McArdle, R., & Beridze, I. (2020). *Malicious Uses and Abuses of Artificial Intelligence*. Trend Micro Research, Europol's European Cybercrime Centre (EC3), & United Nations Interregional Crime and Justice Research Institute (UNICRI). https://www.europol.europa.eu/publications-events/publications/malicious-uses-and-abuses-of-artificial-intelligence

Ciftci, U. A., & Demir, I. (2020). FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1. https://doi.org/10.1109/TPAMI.2020.3009287

Ciftci, U. A., Demir, I., & Yin, L. (2020). How Do the Hearts of Deep Fakes Beat? Deep Fake Source Detection via Interpreting Residuals with Biological Signals. *ArXiv:2008.11363 [Cs]*. http://arxiv.org/abs/2008.11363

Cohen, F. (1987). Computer viruses. *Computers & Security*, *6*(1), 22–35. https://doi.org/10.1016/0167-4048(87)90122-2

Conotter, V., Bodnari, E., Boato, G., & Farid, H. (2014). Physiologically-based detection of computer generated faces in video. *2014 IEEE International Conference on Image Processing (ICIP)*, 248–252. https://doi.org/10.1109/ICIP.2014.7025049

deepfakes. (2022). *Deepfakes_faceswap* [Python]. https://github.com/deepfakes/faceswap (Original work published 2017)

de Lima, O., Franklin, S., Basu, S., Karwoski, B., & George, A. (2020). Deepfake Detection using Spatiotemporal Convolutional Networks. *ArXiv:2006.14749 [Cs, Eess]*. http://arxiv.org/abs/2006.14749

Dhall, S., Dwivedi, A. D., Pal, S. K., & Srivastava, G. (2021). Blockchain-based Framework for Reducing Fake or Vicious News Spread on Social Media/Messaging Platforms. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *21*(1), 8:1-8:33. https://doi.org/10.1145/3467019

Ding, X., Raziei, Z., Larson, E. C., Olinick, E. V., Krueger, P., & Hahsler, M. (2019). Swapped Face Detection using Deep Learning and Subjective Assessment. *ArXiv:1909.04217 [Cs, Stat]*. http://arxiv.org/abs/1909.04217

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). The DeepFake Detection Challenge (DFDC) Dataset. *ArXiv:2006.07397 [Cs]*. http://arxiv.org/abs/2006.07397

Dong, J., & Xie, X. (2021). Visually Maintained Image Disturbance Against Deepfake Face Swapping. *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. doi:10.1109/ICME51207.2021.9428173

Do Nhu, T., Na, I., & Kim, S. H. (2018). *Forensics Face Detection From GANs Using Convolutional Neural Network*.

Durall, R., Keuper, M., Pfreundt, F.-J., & Keuper, J. (2019). *Unmasking DeepFakes with simple Features*. https://doi.org/10.48550/arXiv.1911.00686

England, P., Malvar, H. S., Horvitz, E., Stokes, J. W., Fournet, C., Burke-Aguero, R., Chamayou, A., Clebsch, S., Costa, M., Deutscher, J., Erfani, S., Gaylor, M., Jenks, A., Kane, K., Redmiles, E. M., Shamis, A., Sharma, I., Simmons, J. C., Wenker, S., & Zaman, A. (2021). AMP: Authentication of media via provenance. In *Proceedings of the 12th ACM Multimedia Systems Conference* (pp. 108–121). Association for Computing Machinery. https://doi.org/10.1145/3458305.3459599

Egger, B., Smith, W. A. P., Tewari, A., Wuhrer, S., Zollhoefer, M., Beeler, T., Bernard, F., Bolkart, T., Kortylewski, A., Romdhani, S., Theobalt, C., Blanz, V., & Vetter, T. (2020). 3D Morphable Face Models—Past, Present and Future. *ArXiv:1909.01815 [Cs]*. http://arxiv.org/abs/1909.01815

Facebook AI. (2020, June 25). *Deepfake Detection Challenge Dataset*. https://ai.facebook.com/datasets/dfdc

Fan, L., Li, W., & Cui, X. (2021). Deepfake-Image Anti-Forensics with Adversarial Examples Attacks. *Future Internet*, *13*(11), 288. https://doi.org/10.3390/fi13110288

Fernandes, S., Raj, S., Ewetz, R., Pannu, J. S., Jha, S. K., Ortiz, E., Vintila, I., & Salter, M. (2020). Detecting Deepfake Videos using Attribution-Based Confidence Metric. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. https://doi.org/10.1109/CVPRW50498.2020.00162

Fernando, T., Fookes, C., Denman, S., & Sridharan, S. (2019). Exploiting Human Social Cognition for the Detection of Fake and Fraudulent Faces via Memory Networks. *ArXiv:1911.07844 [Cs, Stat]*. http://arxiv.org/abs/1911.07844

FireEye. (n.d.). *What is a Zero-Day Exploit?* FireEye. Retrieved April 5, 2022, from https://www.fireeye.com/current-threats/what-is-a-zero-day-exploit.html

Fraga-Lamas, P., & Fernández-Caramés, T. M. (2020). Fake News, Disinformation, and Deepfakes: Leveraging Distributed Ledger Technologies and Blockchain to Combat Digital Deception and Counterfeit Reality. *IT Professional*, *22*(2), 53–59. https://doi.org/10.1109/MITP.2020.2977589

Fuzzdb-project/fuzzdb. (2022). [PHP]. FuzzDB Project. https://github.com/fuzzdb-project/fuzzdb (Original work published 2015)

Gabriela Galindo, “XR Belgium posts deepfake of Belgian premier linking Covid-19 with climate crisis,” The Brussels Times, November 9, 2020. https://www.brusselstimes.com/news/belgium-all-news/politics/106320/xr-belgium-posts-deepfake-of-belgian-prem ier-linking-covid-19-with-climate-crisis/

Garrido, P., Valgaerts, L., Sarmadi, H., Steiner, I., Varanasi, K., Pérez, P., & Theobalt, C. (2015). VDub: Modifying Face Video of Actors for Plausible Visual Alignment to a Dubbed Audio Track. *Computer Graphics Forum*, *34*(2), 193–204. https://doi.org/10.1111/cgf.12552

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). *Generative Adversarial Networks*. https://doi.org/10.48550/arXiv.1406.2661

Graphika Team. (2021). *Fake Cluster Boosts Huawei: Accounts with GAN Attack Belgium Over 5G Restrictions*. Graphika. https://public-assets.graphika.com/reports/graphika_report_fake_cluster_boosts_huawei.pdf

Güera, D., & Delp, E. J. (2018). Deepfake Video Detection Using Recurrent Neural Networks. *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 1–6. https://doi.org/10.1109/AVSS.2018.8639163

Handley, M. (2006, July). Why the Internet only just works. *Bt Technology Journal - BT TECHNOL J*, *24*, 119–129. doi:10.1007/s10550-006-0084-z

Hao, H., Baireddy, S., Reibman, A. R., & Delp, E. J. (2020). FaR-GAN for One-Shot Face Reenactment. *ArXiv:2005.06402 [Cs]*. http://arxiv.org/abs/2005.06402

Hartman, T., & Satter, R. (2020, July 15). *These Faces Are Not Real: How to Detect DeepFake Faces*. Reuters. https://graphics.reuters.com/CYBER-DEEPFAKE/ACTIVIST/nmovajgnxpa/

Hasan, H. R., & Salah, K. (2019). Combating Deepfake Videos Using Blockchain and Smart Contracts. *IEEE Access*, *7*, 41596–41606. https://doi.org/10.1109/ACCESS.2019.2905689

HB 198 Election Law - Online Campaign Material - Use of Deepfakes, Ma. (2020). https://legiscan.com/MD/bill/HB198/2020

He, Z., Zuo, W., Kan, M., Shan, S., & Chen, X. (2018). AttGAN: Facial Attribute Editing by Only Changing What You Want. *ArXiv:1711.10678 [Cs, Stat]*. http://arxiv.org/abs/1711.10678

He, Z., Kan, M., Zhang, J., & Shan, S. (2020). PA-GAN: Progressive Attention Generative Adversarial Network for Facial Attribute Editing. *ArXiv:2007.05892 [Cs]*. http://arxiv.org/abs/2007.05892

Hsu, C.-C., Zhuang, Y.-X., & Lee, C.-Y. (2020). Deep Fake Image Detection Based on Pairwise Learning. *Applied Sciences*, *10*(1), 370. https://doi.org/10.3390/app10010370

Huang, R., Zhang, S., Li, T., & He, R. (2017). Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis. *ArXiv:1704.04086 [Cs]*. http://arxiv.org/abs/1704.04086

Huang, X., & Belongie, S. (2017). Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. *ArXiv:1703.06868 [Cs]*. http://arxiv.org/abs/1703.06868

Huang, H., Wang, Y., Chen, Z., Zhang, Y., Li, Y., Tang, Z., Chu, W., Chen, J., Lin, W., & Ma, K.-K. (2021). CMUA-Watermark: A Cross-Model Universal Adversarial

Watermark for Combating Deepfakes. *ArXiv:2105.10872 [Cs]*. http://arxiv.org/abs/2105.10872

Johansen, A. G. (2020, August 13). *How to spot deepfake videos—15 signs to watch for*. Norton. https://us.norton.com/internetsecurity-emerging-threats-how-to-spot-deepfakes.html

Joshi, S. (2021, September 7). *They Follow You on Instagram, Then Use Your Face To Make Deepfake Porn in This Sex Extortion Scam*. https://www.vice.com/en/article/z3x9yj/india-instagram-sextortion-phishing-deepfake-porn-scam

Kaggle. (2020). *Deepfake Detection Challenge*. https://kaggle.com/competitions/deepfake-detection-challenge

Kana, M. (2020, March 28). *Variational Autoencoders (VAEs) for Dummies—Step By Step Tutorial | Towards Data Science*. https://towardsdatascience.com/variational-autoencoders-vaes-for-dummies-step-by-step-tutorial-69e6d1c9d8e9

Kana, M. (2021, February 19). *Generative Adversarial Network (GAN) for Dummies—A Step By Step Tutorial*. Medium. https://towardsdatascience.com/generative-adversarial-network-gan-for-dummies-a-step-by-step-tutorial-fdefff170391

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). Progressive Growing of GANs for Improved Quality, Stability, and Variation. *ArXiv:1710.10196 [Cs, Stat]*. http://arxiv.org/abs/1710.10196

Karras, T., Laine, S., & Aila, T. (2019). *A Style-Based Generator Architecture for Generative Adversarial Networks*. 4401–4410. https://openaccess.thecvf.com/content_CVPR_2019/html/Karras_A_Style-Based_Generator_Architecture_for_Generative_Adversarial_Networks_CVPR_2019_paper.html

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). *Analyzing and Improving the Image Quality of StyleGAN*. 8110–8119. https://openaccess.thecvf.com/content_CVPR_2020/html/Karras_Analyzing_and_Improving_the_Image_Quality_of_StyleGAN_CVPR_2020_paper.html

Keldenich, T. (2021, October 17). Encoder Decoder What and Why ? - Simple Explanation. *Inside Machine Learning*. https://inside-machinelearning.com/en/encoder-decoder-what-and-why-simple-explanation/

Khalid, H., & Woo, S. S. (2020). *OC-FakeDect: Classifying Deepfakes Using One-Class Variational Autoencoder*. 656–657. https://openaccess.thecvf.com/content_CVPRW_2020/html/w39/Khalid_OC-FakeDect_Classifying_Deepfakes_Using_One-Class_Variational_Autoencoder_CVPRW_2020_paper.html

Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Niessner, M., Pérez, P., Richardt, C., Zollhöfer, M., & Theobalt, C. (2018). Deep video portraits. *ACM Transactions on Graphics*, *37*(4), 163:1-163:14. https://doi.org/10.1145/3197517.3201283

Kim, B.-H., & Ganapathi, V. (2019). LumièreNet: Lecture Video Synthesis from Audio. *ArXiv:1907.02253 [Cs, Eess, Stat]*. http://arxiv.org/abs/1907.02253

Korshunov, P., & Marcel, S. (2018). Speaker Inconsistency Detection in Tampered Video. *2018 26th European Signal Processing Conference (EUSIPCO)*, 2375–2379. https://doi.org/10.23919/EUSIPCO.2018.8553270

Korshunov, P., Halstead, M., Castan, D., Graciarena, M., McLaren, M., Burns, B., & Marcel, S. (2019, June). Tampered speaker inconsistency detection with phonetically aware audio-visual features. In *International Conference on Machine Learning*.

Koopman, M., Macarulla Rodriguez, A., & Geradts, Z. (2018, August 20). *Detection of Deepfake Video Manipulation*.

Le, T.-N., Nguyen, H. H., Yamagishi, J., & Echizen, I. (2022). *Robust Deepfake On Unrestricted Media: Generation And Detection*. doi:10.48550/ARXIV.2202.06228

Li, J., Zhao, B., & Zhang, C. (2018). Fuzzing: A survey. *Cybersecurity*, *1*(1), 6. https://doi.org/10.1186/s42400-018-0002-y

Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., & Guo, B. (2020). Face X-ray for More General Face Forgery Detection. *ArXiv:1912.13458 [Cs]*. http://arxiv.org/abs/1912.13458

Li, Y., Chang, M.-C., & Lyu, S. (2018). In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, 1–7. https://doi.org/10.1109/WIFS.2018.8630787

Li, Y., & Lyu, S. (2019). Exposing DeepFake Videos By Detecting Face Warping Artifacts. *ArXiv:1811.00656 [Cs]*. http://arxiv.org/abs/1811.00656

Lima, C. (2021, August 6). Analysis | The Technology 202: As senators zero in on deepfakes, some experts fear their focus is misplaced. *Washington Post*.

https://www.washingtonpost.com/politics/2021/08/06/technology-202-senators-zero-deepfakes-some-experts-fear-their-focus-is-misplaced/

Liu, M., Ding, Y., Xia, M., Liu, X., Ding, E., Zuo, W., & Wen, S. (2019). STGAN: A Unified Selective Transfer Network for Arbitrary Image Attribute Editing. *ArXiv:1904.09709 [Cs]*. http://arxiv.org/abs/1904.09709

Luan, X., Geng, H., Liu, L., Li, W., Zhao, Y., & Ren, M. (2020). Geometry Structure Preserving Based GAN for Multi-Pose Face Frontalization and Recognition. *IEEE Access*, *8*, 104676–104687. https://doi.org/10.1109/ACCESS.2020.2996637

Nagano, K., Seo, J., Xing, J., Wei, L., Li, Z., Saito, S., Agarwal, A., Fursund, J., & Li, H. (2018). paGAN: Real-time avatars using dynamic textures. *ACM Transactions on Graphics*, *37*(6), 258:1-258:12. https://doi.org/10.1145/3272127.3275075

Natsume, R., Yatagawa, T., & Morishima, S. (2018). RSGAN: Face Swapping and Editing using Face and Hair Representation in Latent Spaces. *ArXiv:1804.03447 [Cs]*. http://arxiv.org/abs/1804.03447

Natsume, R., Yatagawa, T., & Morishima, S. (2018). FSNet: An Identity-Aware Generative Model for Image-based Face Swapping. *ArXiv:1811.12666 [Cs]*. http://arxiv.org/abs/1811.12666

Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(8), e2120481119. https://doi.org/10.1073/pnas.2120481119

Nimmo, B., Eib, C. S., Tamora, L., Johnson, K., Smith, I., Buziashvili, E., Kann, A., Karan, K., Ponce de León Rosas, E., & Rizzuto, M. (2019). *#OperationFFS: Fake Face*

Swarm. Graphika, DFRLab. https://public-assets.graphika.com/reports/graphika_report_operation_ffs_fake_face_storm.pdf

Nirkin, Y., Keller, Y., & Hassner, T. (2019). *FSGAN: Subject Agnostic Face Swapping and Reenactment*. 7184–7193. https://openaccess.thecvf.com/content_ICCV_2019/html/Nirkin_FSGAN_Subject_Agnostic_Face_Swapping_and_Reenactment_ICCV_2019_paper.html

Nirkin, Y., Masi, I., Tran, A. T., Hassner, T., & Medioni, G. (2017). On Face Segmentation, Face Swapping, and Face Perception. *ArXiv:1704.06729 [Cs]*. http://arxiv.org/abs/1704.06729

Nirkin, Y., Wolf, L., Keller, Y., & Hassner, T. (2020). DeepFake Detection Based on the Discrepancy Between the Face and its Context. *ArXiv:2008.12262 [Cs]*. http://arxiv.org/abs/2008.12262

Ma, T., Li, D., Wang, W., & Dong, J. (2021). CFA-Net: Controllable Face Anonymization Network with Identity Representation Manipulation. *ArXiv:2105.11137 [Cs]*. http://arxiv.org/abs/2105.11137

Marra, F., Gragnaniello, D., Verdoliva, L., & Poggi, G. (2018). *Do GANs leave artificial fingerprints?* https://doi.org/10.48550/arXiv.1812.11842

Masood, M., Nawaz, M., Malik, K. M., Javed, A., & Irtaza, A. (2021). *Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward*. 54.

Matern, F., Riess, C., & Stamminger, M. (2019). Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations. *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 83-92.

Menéndez, H. D., Clark, D., & T. Barr, E. (2021). Getting Ahead of the Arms Race: Hothousing the Coevolution of VirusTotal with a Packer. *Entropy*, *23*(4), 395. https://doi.org/10.3390/e23040395

Mirsky, Y., & Lee, W. (2022). The Creation and Detection of Deepfakes: A Survey. *ACM Computing Surveys*, *54*(1), 1–41. https://doi.org/10.1145/3425780

MITRE. (2021, October). *CAPEC - CAPEC-28: Fuzzing (Version 3.7)* [MITRE]. https://capec.mitre.org/data/definitions/28.html

Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020). Emotions Don't Lie: An Audio-Visual Deepfake Detection Method Using Affective Cues. *ArXiv:2003.06711 [Cs]*. http://arxiv.org/abs/2003.06711

Mo, H., Chen, B., & Luo, W. (2018). Fake Faces Identification via Convolutional Neural Network. *Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security*, 43–47. https://doi.org/10.1145/3206004.3206009

Paganini, P. (2020, August). *The most common social engineering attacks [updated 2020]*. Infosec Resources. https://resources.infosecinstitute.com/topic/common-social-engineering-attacks/

Perov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S., Umé, C., Dpfks, M., Facenheim, C. S., RP, L., Jiang, J., Zhang, S., Wu, P., Zhou, B., & Zhang, W. (2021). DeepFaceLab: Integrated, flexible and extensible face-swapping framework. *ArXiv:2005.05535 [Cs, Eess]*. http://arxiv.org/abs/2005.05535

Prabhu, A., Materzyńska, J., Dokania, P. K., Torr, P. H. S., & Lim, S.-N. (2020, June 15). *Is Your Face Fake? Social Impacts of Algorithmic "Fake" Determination* [Conference]. DFDC Risk-a-thon & CVPR media forensics workshop 2020, Seattle Washington. https://drimpossible.github.io/documents/dfdc_slides.pdf

Pu, J., Mangaokar, N., Kelly, L., Bhattacharya, P., Sundaram, K., Javed, M., Wang, B., & Viswanath, B. (2021). *Deepfake Videos in the Wild: Analysis and Detection*. https://arxiv-org.proxy01.its.virginia.edu/abs/2103.04263v2

Pumarola, A., Agudo, A., Martinez, A. M., Sanfeliu, A., & Moreno-Noguer, F. (2018). GANimation: Anatomically-aware Facial Animation from a Single Image. *ArXiv:1807.09251 [Cs]*. http://arxiv.org/abs/1807.09251

Qian, S., Lin, K.-Y., Wu, W., Liu, Y., Wang, Q., Shen, F., Qian, C., & He, R. (2019). Make a Face: Towards Arbitrary High Fidelity Face Manipulation. *ArXiv:1908.07191 [Cs]*. http://arxiv.org/abs/1908.07191

Rana, M., & Sung, A. (2020). *DeepfakeStack: A Deep Ensemble-based Learning Technique for Deepfake Detection*. 70–75. https://doi.org/10.1109/CSCloud-EdgeCom49738.2020.00021

Rothkopf, J. (2020, July 1). Deepfake Technology Enters the Documentary World. *The New York Times*. https://www.nytimes.com/2020/07/01/movies/deepfakes-documentary-welcome-to-chechnya.html

Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, W., Masi, I., & Natarajan, P. (2019). Recurrent Convolutional Strategies for Face Manipulation Detection in Videos. *ArXiv:1905.00582 [Cs]*. http://arxiv.org/abs/1905.00582

Sanchez, E., & Valstar, M. (2018). Triple consistency loss for pairing distributions in GAN-based face synthesis. *ArXiv:1811.03492 [Cs]*. http://arxiv.org/abs/1811.03492

SB No. 751 An Act Relating to the Creation of a Criminal Offense For Fabricating a Deceptive Video with Intent to Influence the Outcome of an Election. Tx Section 255.004 (2019). https://capitol.texas.gov/tlodocs/86R/billtext/html/SB00751F.htm

SB 1988 An Act To Prohibit the Distribution of Deceptive Images or Audio or Video Recordings with the Intent To Influence the Outcome of an Election. Me. (2019). https://legiscan.com/ME/text/LD1988/2019

SB 6513 Restricting the use of deepfake audio or visual media in campaigns for elective office.Wa (2020). https://apps.leg.wa.gov/billsummary/?BillNumber=6513&Year=2020&Initiative=false

Schwartz, O. (2018, November 12). You thought fake news was bad? Deep fakes are where truth goes to die. *The Guardian*. https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth

Seferbekov, S. (2022). *Selimsef/dfdc_deepfake_challenge* [Python]. https://github.com/selimsef/dfdc_deepfake_challenge (Original work published 2020)

Segalis, E., & Galili, E. (2020). OGAN: Disrupting Deepfakes with an Adversarial Attack that Survives Training. *ArXiv:2006.12247 [Cs, Stat]*. http://arxiv.org/abs/2006.12247

shaoanlu. (2022). *Faceswap-GAN* [Jupyter Notebook]. https://github.com/shaoanlu/faceswap-GAN (Original work published 2017)

Yujun Shen, Ping Luo, Junjie Yan, Xiaogang Wang, and Xiaoou Tang. 2018. FaceID-GAN: Learning a symmetry three-player GAN for identity-preserving face synthesis. In

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 821–830.

Yujun Shen, Bolei Zhou, Ping Luo, and Xiaoou Tang. 2018. FaceFeat-GAN: A two-stage approach for identitypreserving face synthesis. arXiv preprint arXiv:1812.01288 (2018).

Sherman, I. N., Stokes, J. W., & Redmiles, E. M. (2021). Designing Media Provenance Indicators to Combat Fake Media. *24th International Symposium on Research in Attacks, Intrusions and Defenses*, 324–339. https://doi.org/10.1145/3471621.3471860

smartcontract694. (2021). *Smartcontract694/PoA*. https://github.com/smartcontract694/PoA (Original work published 2018)

Stouffer, C. (2021, September 15). *What is scareware? A definition, examples, removal tips*. https://us.norton.com/internetsecurity-online-scams-how-to-spot-online-scareware-scams.html

Straub, J. (2019). Using subject face brightness assessment to detect 'deep fakes' (Conference Presentation). In N. Kehtarnavaz & M. F. Carlsohn (Eds.), *Real-Time Image Processing and Deep Learning 2019* (p. 18). SPIE. https://doi.org/10.1117/12.2520546

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic Attribution for Deep Networks. *ArXiv:1703.01365 [Cs]*. http://arxiv.org/abs/1703.01365

Tan, M., & Le, Q. V. (2020). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *ArXiv:1905.11946 [Cs, Stat]*. http://arxiv.org/abs/1905.11946

Tariq, S., Lee, S., Kim, H., Shin, Y., & Woo, S. S. (2018). Detecting Both Machine and Human Created Fake Face Images In the Wild. *MPS@CCS*. https://doi.org/10.1145/3267357.3267367

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. *Engineering Applications of Artificial Intelligence*, *110*, 104673. https://doi.org/10.1016/j.engappai.2022.104673

Twitter. (n.d.). *Our synthetic and manipulated media policy | Twitter Help*. Retrieved March 17, 2022, from https://help.twitter.com/en/rules-and-policies/manipulated-media

Vougioukas, K., Petridis, S., & Pantic, M. (2019). *End-to-End Speech-Driven Realistic Facial Animation with Temporal GANs*. 4.

Wang, R., Juefei-Xu, F., Ma, L., Xie, X., Huang, Y., Wang, J., & Liu, Y. (2020). FakeSpotter: A Simple yet Robust Baseline for Spotting AI-Synthesized Fake Faces. *ArXiv:1909.06122 [Cs]*. http://arxiv.org/abs/1909.06122

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., & Catanzaro, B. (2018). High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. *ArXiv:1711.11585 [Cs]*. http://arxiv.org/abs/1711.11585

Westerlund, M. (2019). The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review*, *9*(11), 40–53. https://doi.org/10.22215/timreview/1282

Wu, W., Zhang, Y., Li, C., Qian, C., & Loy, C. C. (2018). ReenactGAN: Learning to Reenact Faces via Boundary Transfer. *ArXiv:1807.11079 [Cs]*. http://arxiv.org/abs/1807.11079

Xuan, X., Peng, B., Wang, W., & Dong, J. (2019). On the generalization of GAN image forensics. https://doi.org/10.48550/arXiv.1902.11153

Yeh, C.-Y., Chen, H.-W., Tsai, S.-L., & Wang, S.-D. (2020). *Disrupting Image-Translation-Based DeepFake Algorithms with Adversarial Attacks*. 53–62. https://openaccess.thecvf.com/content_WACVW_2020/html/w4/Yeh_Disrupting_Image-Translation-Based_DeepFake_Algorithms_with_Adversarial_Attacks_WACVW_2020_paper.html

Yang, X., Li, Y., & Lyu, S. (2018). *Exposing Deep Fakes Using Inconsistent Head Poses*. https://doi.org/10.48550/arXiv.1811.00661

Youtube. (n.d.). *Misinformation policies—YouTube Help*. Retrieved March 17, 2022, from https://support.google.com/youtube/answer/10834785?hl=en

Zakharov, E., Shysheya, A., Burkov, E., & Lempitsky, V. (2019). Few-Shot Adversarial Learning of Realistic Neural Talking Head Models. *ArXiv:1905.08233 [Cs]*. http://arxiv.org/abs/1905.08233

Zhang, G., Kan, M., Shan, S., & Chen, X. (2018). *Generative Adversarial Network with Spatial Attention for Face Attribute Editing*. 417–432. https://openaccess.thecvf.com/content_ECCV_2018/html/Gang_Zhang_Generative_Adversarial_Network_ECCV_2018_paper.html

Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2019). Self-Attention Generative Adversarial Networks. *Proceedings of the 36th International Conference on Machine Learning*, 7354–7363. https://proceedings.mlr.press/v97/zhang19d.html

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. (2017). StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks. *ArXiv:1612.03242 [Cs, Stat]*. http://arxiv.org/abs/1612.03242

Zhang, Y., Zhang, S., He, Y., Li, C., Loy, C. C., & Liu, Z. (2019). One-shot Face Reenactment. *ArXiv:1908.03251 [Cs, Eess]*. http://arxiv.org/abs/1908.03251

Zhang, Y., Zheng, L., & Thing, V. L. L. (2017). Automated face swapping and its detection. *2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP)*, 15–19. https://doi.org/10.1109/SIPROCESS.2017.8124497