GAMES OF HIDE AND SEEK:
How the Malware Arms Race Informs Analysis on the Future of DeepFake Generation and Mitigation

a bit wordy – gets the right things across, but need to shorten.

Abstract:

no need to underline or :

On Wednesday March 16th, 2022, a group of hackers broadcast a fake video of Ukranian 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 the rise in popularity of deep-learning based face-swaps on Reddit in 2017 (Aubé, 2017). The quality of the Zelenskyy DeepFake was not state ~~of the~~ art and contained visual and auditory artifacts that allowed users to easily identify the video as fake. However, more sophisticated ~~generations of~~ DeepFakes are not ~~as~~ easily distinguishable from authentic content. In order to engage in academic discourse on the contemporary threat of DeepFakes, the is ~~technical paper reviews academic literature on state-of-the-art~~ technology to provide s an overview of DeepFake generation methods, ~~an overview~~ of DeepFake countermeasures, and an ~~overview~~ of the current performance of DeepFake generation and DeepFake 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 the DeepFake generation and mitigation arms race.

← really limited to Reddit?

avoid repetition

summarizes surveys

1. Overview of 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. 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 (Luan et al., 2020). Facial swapping via computer vision has been proposed and used 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. These DeepFakes have garnered significant attention as an attack surface.

odd use of this

use section numbers to make organization more clear
?

1. 1 Kinds

~~Categories~~ 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). Face-swaps involve replacing the face of a subject in a source video with the identity of a target subject. Reenactment DeepFakes involve using footage of a source subject to drive the expressions, pose, gaze or body movements of a target subject. Lip-syncing 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?

~~Overview of~~ DeepFake Attack Surfaces~~:~~

Reenactment and replacement DeepFakes have gotten the most attention from research and political communities, due to threatening attack surfaces that undermine trust in truth. Puppetry and face swap methods can be used to target individuals in defamation and discredibility attacks. In 2018, journalist Rana Ayyub was targeted by political opponents in a defamation attack that spread a pornographic DeepFake video featuring her face. The video was shared so many times and resulted in such a degree of cyber harassment that the United Nations intervened to call for Ayyub's protection (Ayyub, 2018). Additionally, face-swap and reenactment applications can be used to tamper with evidence. The application of face-swap and reenactment videos to target political leaders in disinformation attacks has gained the attention of law-makers. Laws passed in Texas and California ban the distribution of deceptive audio or visual media around elections (8 CA AB 730; TX SB 751). Similar legislation has been introduced in Maine, Washington and Maryland (HB 198, 2020 Regular Sess. (Md. 2020); LD 1988, 129th Legislature, (Maine 2020); SB 6513, 2020 Regular Sess. (Wash. 2020)). Additionally, disinformation attacks have been used in the wild. Ukrainian President Volodymyr Zelenskyy was targeted in an attack that involved a face-swap video that was broadcast on live television. In this forgery, the Ukranian President was depicted calling on soldiers to lay down their weapons in the war against Russia (Allyn, 2022). Mirsky & Lee describe deception by child predators as an attack surface for deep learning based facial editing (Mirsky & Lee, 2022). They propose that predators can edit photos to appear younger. Face synthesis models have seen applications in the generation of fake online profiles. These profiles have been used 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). DeepFakes have seen increasing application in the wild as a social engineering tool in phishing schemes. Phishing attacks use auditory DeepFakes

repetitive w/ Intro (if intentional, add "As mentioned in the Intro, ...")

as a vector for impersonation in voice phishing (Brewster, 2021). Reenactment DeepFakes, such as pornographic face-swaps, have been used as a blackmail tool in other extortion schemes (Joshi, 2021).

1.3? Generating Deep Fakes

~~Machine Learning Foundations: Deep Neural Networks:~~

DeepFake generation is driven by deep neural networks (DNN), a ~~specific~~ type of machine learning algorithm that makes decisions by passing data through multiple layers of statistical models ~~designed~~ trained to learn underlying patterns in the data. These layers are modeled after neural connections in the brain and include "hidden" layers that sort, process and order data based on different features. Central components of a neural network are the "synapses" of each layer, the activation function and objective loss. The "synapses" hold information on the concepts learned through the model by storing an associated weight. The activation function summarizes a neuron's output. Objective loss is used to evaluate the performance of a given set of model weights. The output of the network is computed through a process of forward propagation by which input is passed through the layers of the neural network and interpreted by each layer using the activation function. The model is optimized through a process of backpropagation where the performance is evaluated using the objective loss function and weights are tweaked to minimize error (Bonner, 2019). Different DeepFake generators use different neural network architectures and mechanisms to manipulate, replicate, and fine-tune image and sound data. Commonly deployed forms of DNN's for DeepFake generation are autoencoders, generative adversarial networks, encoder-decoder networks, recurrent neural networks, and variational autoencoders (Mirsky & Lee, 2021).

don't think this will be usefully understood by readers who are not already familiar with DNNs

~~DeepFake Model Architecture:~~

diagrams would help

Encoder-Decoder Networks:

Early face swap models were built using two encoder-decoder network pairs (deepfakes, 2017; Masood et al., 2021). Often used in translation applications, an encoder-decoder network 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).

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. The RSGAN architecture proposed by Natsume et al. uses two variational autoencoders to target training towards facial and hair regions separately (Natsume et al., 2018a). The FSNET architecture proposed by the same authors targets training towards the face region in source images and non-face regions in target images (Natsume et al., 2018b). The FSNET architecture also uses a process of inpainting in the generator. Bao et al. propose a CVAE_GAN model that combin1`es a generative adversarial network and variational autoencoder to condition generation on fine-grained categories. The authors propose image inpainting as one task that this architecture is well-suited for (Bao et al., 2017). Building off of work on the CVAE_GAN, Qian et al. propose an Additive Focal Variational Auto-encoder model that targets training towards appearance encodings and identity-agnostic expression encodings (Qian et al., 2019). Other applications of variational autoencoders include Kim & Ganapathi's lip-sync model which drives facial expressions, body postures and gestures with source audio (Kim & Ganapathi, 2019).

Variational autoencoder architectures improve the performance of DeepFake generation for source data with inappropriate fitting for 3D morphable models (ie. strange lighting conditions, 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).

3D Morphable Models:

if available, example images would be useful

In 1999, Blanz and Vetter proposed a 3D morphable face model that allowed for the generation of 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). The reception of this paper was such that it received an ? impact award and has recently seen a rise in deep learning applications (Egger et al., 2020). Kim et al. leverage 3D morphable models in an architecture that can be utilized in DeepFake puppetry applications that drive full head animation rather than only facial expressions (Kim et. al, 2018). Shen et al. apply a 3D morphable model in tandem with an encoder-decoder network to produce facial reenactment with reduced identity leakage (Shen et. al, 2018a). These authors later proposed an improved model that leveraged 3 encoder-predictor networks and a set of GANs to produce reenactments. One of these encoder-predictor networks was trained to predict 3D morphable model parameters (Shen et al., 2018b). Nagano et. al utilize 3D morphable models to generate 3D deepfakes (Nagano et al, 2018). However, the utility of a 3D visual

this is too much of a list, and not enough of a story forgery has been questioned (Mirsky & Lee, 2021). Nirkin et al. use a fully convolutional neural network to segment facial regions, then apply a 3D morphable model to understand facial texture and geometry (Nirkin et al., 2017). Use cases for 3D facial models have also been seen in lip-sync applications (Garrido et al., 2015).

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 gained popularity in the production of DeepFakes. Faceswap-GAN adds a discriminator and adversarial loss to the encoder-decoder network popularized by reddit user deepfakes (shaoanlu, 2017). In 2019, Nirkin et. al proposed a subject agnostic face-swapping and reenactment model built using generative adversarial networks. This model implemented a network for face completion to handle weaknesses related to facial occlusions and a face blending network to reduce artifacts, while preserving target lighting (Nirkin et al., 2017). Vougioukas et al. applied generative adversarial networks to create a lip-syncing model that uses audio to drive the motions of a talking head. The authors used discriminators to improve audio-visual synchronization, frame quality and video quality (Vougioukas et al., 2019).

Generative adversarial networks have also been applied to face synthesis tasks (Masood et al., 2021). Huang et al's face synthesis model is capable of handling minimal, poorly posed input images. This model leverages a two-pathway generative adversarial network (Huang et al., 2017). Another face synthesis model uses a generative adversarial network to synthesize faces with increased semantic fidelity by handling global dependencies in a self-attention module (Zhang et al., 2017).

A notable advance in GAN training was proposed by Karras et al. in 2017. Their ProGAN methodology utilizes a mini-batch size and additional network layers during training 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 such as gaze direction and teeth alignment (Karras et al., 2020). A number of other researchers have used varying strategies to address the resolution of GAN generated images (Zhang et al., 2019; Brock et al., 2019).

Research has proposed models of generative adversarial networks that translate images across domains to allow for facial attribute manipulation (He et al., 2018; Zhang et al., 2018; Liu et al., 2019; Choi et al., 2020; He et al., 2020). In the domain of reenactment DeepFakes, the Pix2pixHD architecture has proved to produce high resolution images (Masood et. al, 2021).The pix2pixHD architecture uses conditional generative adversarial networks with perceptual loss, a measure of high level differences in images (Wang et al., 2018). Other GAN based approaches have been applied to facial reenactment DeepFakes (Wu et al., 2018; Pumarola et al., 2018; Sanchez & Valstar, 2018). The model proposed by Pumarola et al. conditions generation on annotated Action Units, which encode facial expressions (Pumarola et al., 2018). One weakness of the models proposed by Wu et al., Pumarola et al., Snachez & Valstar and Wang is a reliance on a large corpus of high fidelity training data.(Masood et al., 2021). Facial reenactment models that require less subject-specific training data address this weakness (Zakharov et al., 2019; Zhang et al., 2019b; Hao et al., 2020). Weaknesses of these few-shot learning methods include identity leakage (Masood et al., 2021).

good

?

General Advances in DeepFake Generation:

Recent?

Other notable advances include use of post-processing such as inblending, smoothing and models that address occlusions. These post processing steps address artifacts in generated DeepFakes. Perceptual loss based on the convolutional neural network based VGG-Face vision model is used to improve the fidelity of eye movements and to smooth artifacts (Masood et. al., 2021). Artifacts have also been addressed through the use of loss functions to handle specific weaknesses (Masood et al., 2021). Image fidelity has been improved through variational autoencoder feature disentanglement, self-attention modules, and adaptive instance normalization (AdaIN) layers (Huang & Belongie, 2017; Masood et al, 2021). Temporal coherence has been addressed through optical flow estimation and temporal discriminators (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). One and few-shot models also lower the burden of training data. 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 have also produced more real time manipulations, allowing for a greater enmeshment with other social engineering pressures (Masood et al, 2021).

Not clear why this is separate section, not fitting in with others

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. Generalization refers to a necessity for high-quality training data that matches the identity or identities of the DeepFake subject(s). 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. Identity leakage references to weakness in face swap and reenactment

? not usually what generalization means in ML - are you sure this is what they mean?

okay, used this earlier without explaining. Maybe this belongs earlier (whole section

DeepFakes in which the source identity is reproduced along with the target identity in the generated DeepFake. 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. Temporal coherence refers to artifacts such as flicker and jitter that appear when DeepFake generators operate on a frame by frame basis (Mirsky & Lee, 2022). 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 can result in semantic inconsistencies in generated media. 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).

2. ~~Overview of~~ DeepFake Countermeasures:

A number of varying 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.

Technical Detection Approaches:

Blending Artifact Detection:

Technical approaches for detecting DeepFake media use a number of different strategies. Many DeepFakes contain visual, semantic artifacts, such as discoloration, inconsistent lighting, unnatural teeth, or unnatural hair that indicate to a viewer inauthenticity (Norton, 2020). 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 and neighboring pixels are dissimilar. 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 et al., 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). Mo et al. propose a model that passes an input image through a high pass filter to obtain residuals and uses a convolutional neural network to classify real and fake images based on statistical artifacts (Mo et al., 2018). Li et al. train a convo-lutional neural network explicitly on blending boundaries in order to classify real and fake images (Li et al., 2020). Other researchers have proposed a model to classify real and fake images based on residuals leftover from face-warping processes (Li & Lyu, 2019). The model proposed by Li & Lyu achieves an accuracy between 84% and 99% in experimental testing, however, the authors note limitations of this model toward DeepFakes with higher quality and resolution (Li & Lyu, 2019).

'images available?

Environmental Artifact Detection:

Other artifacts in face-swap DeepFakes are found in 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).

Forensic Artifact Detection:

Targeting yet another semantic inconsistency, Yang et al. propose a network that monitors and predicts facial landmarks to identify DeepFakes based on inconsistent head poses (Yang et al., 2018). Forensics analysis of manipulated media has revealed that each GAN leaves a unique fingerprint in its generated images, allowing for the identification of media's source GAN (Marra et al., 2018). Additionally, each camera produces a unique sensor noise known as Photo Response Non-Uniformity (PNRU). Koopman et al. leverage PNRU analysis in their DeepFake detection model (Koopman 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). Mitall et al. analyze similarity in audio and visual emotional cues to identify digitally forged content (Mittall et al., 2020). Other scholars have proposed models that classify DeepFakes based on the synchronization of audio phonemes and visual mouth shapes (Korshunov & Marcel, 2018; Korshunov et al., 2019).

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 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 have leveraged classifiers that evaluate the temporal coherence of input media. Güera & Delp train a recurrent neural network to catch flicker and jitter indicative of DeepFake videos (Güera & Delp, 2018). Sabir et al. utilize combinations of recurrent convolutional models to exploit temporal information in the classification of DeepFake video content (Sabir et al., 2019). The model proposed by Chan et al. compares sequential frames to classify media content (Chan et al., 2019). Amerini et al. analyze dissimilarity between video frames in their DeepFake classification model (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 train a reconstruction variational autoencoder network solely on real faces. Then, they compute an anomaly score through the mean square error of the encoded and reconstructed images (Khalid & Woo, 2020). Wang et al. monitor the layer-by-layer activation of facial recognition models to differentiate fake and real content (Wang et al., 2020). Fernandes et al. measure how well an image fits the training distribution of the popular VGGFace recognition model (Fernandes et al., 2020).

Generic Classifiers:

Unsupervised deep learning architectures are also deployed in generic classification models trained to differentiate authentic and DeepFake content. 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).

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 and provide more challenge to 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). The research community has noted limits in DeepFake detection generalizability 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

this is very comprehensive & well done. If you can organize into a table or some structure.

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. find that the CapsuleForensics detection model had the highest accuracy for Black faces with an F1 score of 74%. The performance on Caucasian faces is comparable with an F1 score of 72%. However, the performance on Asian faces dropped 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. Detection based on blending artifacts is likely to misclassify the faces of individuals who have conditions such as leprosy or vitiligo. Blending artifact based detection is also likely to misclassify the faces of burn victims, individuals with facial tattoos and smooth baby faces (Prabhu, 2020).

(does this mean it detects DFs with diff. accury based on race of target/ source face ?

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 of applying adversarial loss to images such that manipulating these images is made 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).

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 immutable, tamperproof 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).England et. al. propose an alternative authentication of media via provenance framework, 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).

where does Detection section end?

Content Moderation Approaches:

Other applications of distributed ledger technologies to combat deceptive content have been proposed. Frameworks include the application of Blockchain for decentralized content moderation, trustworthiness checkers, incentivized fact checking, decentralized social media platforms, and reputation systems. Trustworthiness checkers allow any node 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). The use of these approaches requires digital literacy among users.

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, audio and visual evidence can be more critically consumed. Awareness of potentially distinguishable visual artifacts in less sophisticated DeepFake generations allows users to identify some forged content. Greater digital literacy also involves awareness of heuristics to evaluate the validity of information sources and content.

Current State of DeepFake Detection and Generation Arms Race:

Kaggle DeepFake Detection Challenge 2020:

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 developed a novel dataset containing more than 100,000 videos that was used 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. Face swaps were created using convolutional autoencoders, a frame-based morphable mask model, a GAN-based neural talking head model, Nirkin et al.'s FSGAN model and Karras et al's Style-GAN (Nirkin et al, 2019; Karras et al., 2018). Additionally, 70% of videos in this dataset were manipulated with augmentations, such as Gaussian blurring, illumination changes, frame rate alterations, gray-scale conversions, resolution changes and rotations. 30% of videos contained object overlays that served as adversarial distractions (Dolhansky et al., 2020).

Research Commentary On Generation Weaknesses:

Dolhansky et al note that convolutional autoencoders performed best and had the most flexibility. However, this architecture had weaknesses around extreme poses and glasses. Frame-based morphable mask models tend to work well on single-frame images, but produce discontinuities in the face and occasionally fail to fit the mask to a face. The FSGAN model functioned well in good lighting conditions and translated extreme poses well. However, it experienced poor performance in dark lighting conditions. The GAN based neural talking head model performed consistently, but performed poorly in poor lighting conditions and produced visually similar eyes on all DeepFake generations. The StyleGAN model performed the worst of all selected generation methods. Weaknesses of this model included semantically invalid eye poses, such as eyes looking in different directions and mismatched illumination (Dolhansky et al., 2020).

can you
present
summary
numbers?

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 65.18% 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 82.56% 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 predicting a probability of 0.50 on every video in the dataset. 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).

really
4 digits
of
precision?

?

comparing
to random
guessing
for
balanced
test set?

DeepFake Videos In The Wild: Analysis and Detection (2021):

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 models on a DeepFake video test set created from a collection of non-pornographic DeepFakes found on online platforms such as Youtube, Billibilli and Reddit. The videos collected for this test set were contextually presented as DeepFakes (ie. appeared in search results for targeted searches or were published under a DeepFake subforum). 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 in the wild DeepFakes. 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, detection models assume that every frame of a video has a fake face. In the wild, this was not found to be true. Additionally, detection models assume that there is only one face in each frame. DeepFakes in the wild were found to contain multiple faces in a frame. DeepFakes in the wild also 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, Li et. al proposed a Poisson noise DeepFool model that iteratively developed adversarial examples that weakened DeepFake detection accuracy from 0.9997 to 0.0731 (Li et. al, 2021).

Proliferation of DeepFake Profile Photos:

relevance? seems out of place here

Recently, Stanford researchers uncovered more than 1,000 fake LinkedIn profiles that utilize GAN forged profile pictures. These accounts were used to circumvent LinkedIn limits on sales messaging (Bond, 2022). A network of 14 fake Twitter accounts, with GAN generated profile images, was discovered by Graphika in late 2020. These accounts were used to amplify anti-Belgian content (Graphika, 2021). In 2019, Facebook removed a network of over 900 pages, accounts, and groups tied to large-scale coordinated, inauthentic behavior intended to direct users towards Pro-Trump content before the 2020 election. In this network, dozens of accounts used DeepFake profile pictures (Graphika, 2019).

Games of Hide And Seek:

While the DeepFake generation and detection arms race poses unique challenges, it is not the first situation where the technology of malicious adversaries has been paired against complementary technology developed to combat it.

A History of Malware Generation and Anti-Malware Detection:

The arms race 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 added additional

diagram?

overhead to the process of signature scanning, since it was 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. 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. 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).

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, verifying that programs do contain malicious code necessitates significant computational overhead. Perfect anti-virus would require proof that every executable program on a computer does not contain malicious code. Even if theoretically perfect signature scanners were possible, the computational burden of this task is infeasible for modern computers. 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).

disconnect

sig scanning is very efficient

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. Perfect DeepFake detection would require flexible identification of any falsified media, which is a much broader domain. Additionally, catching all instances of DeepFakes would require analyzing all media uploaded to the internet. Digital platforms, such as Youtube and Billibilli, have the power to enforce constraints on uploads, but the money and resources necessary to implement DeepFake detection is significant. Top performing DeepFake detection algorithms require access to graphical processing units and sufficient memory (Hao, 2020; Seferbekov, 2020). Given that access to GPU is not universal for local machines, the resources necessary to analyze videos for fabricated content would need to be hosted by digital service providers. On average 500 hours of video footage per minute are uploaded to Youtube (Bernaciak & Ross, 2022). Scaling the image processing computations to the number of videos uploaded to the internet daily presents a significant need for computational resources. Given the computational burden of proposed methods of media authentication, ubiquitous deployment of research methodologies to all uploaded videos is unlikely. Many social media platforms have policies around misinformation including DeepFakes. However, it is not clear how DeepFakes are identified in the content moderation procedures of these businesses or whether technical detection algorithms are applied to flagged content (Bickert, 2020; Twitter, n.d; Youtube, n.d.). Though there is significant political pressure to reduce the spread of misinformation on technology platforms, it does not seem likely that deep learning detection algorithms will see practical use in content distribution beyond potential corner case use as a data point in a more complicated user-initiated content moderation procedure.

don't buy this - they scan all content (youtube) for copyright issues

only matters to catch the ones that are viewed

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 internet protocols are a historically slow process, due to contention over benefits, trade-offs and backwards compatibility. [cite source for this]. 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.

impact on privacy?

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 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.

why hide & seek?

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.

Semantic Context and Its Effect on Model Training Decisions:

Not all DeepFake attack surfaces are created equally or warrant the same treatment. A 2019 web crawl by DeepTrace found 15,000 deepfake videos published online. Of these 15,000, 96% of them were pornographic (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.

wow - was this really a "fair" sample, or were they looking for it?

Conclusion?

References