Seeing is Believing: Exploring Perceptual Differences in DeepFake Videos

Rashid Tahir
University of Prince Mugrin
Madina, Saudi Arabia
r.tahir@upm.edu.sa

Brishna Batool
Hira Jamshed
Mahnoor Jameel
Mubashir Anwar
Lahore University of Management Sciences
Lahore, Pakistan

Faizan Ahmed
University of Virginia
Charlottesville, USA
faizan288@gmail.com

Muhammad Adeel Zaffar
Suleman Dawood School of Business, Lahore University of Management Sciences
Lahore, Pakistan
adeel.zaffar@lums.edu.pk

Muhammad Fareed Zaffar
Lahore University of Management Sciences
Lahore, Pakistan
fareed.zaffar@lums.edu.pk

ABSTRACT

With AI on the boom, DeepFakes have emerged as a tool with a massive potential for abuse. The hyper-realistic imagery of these manipulated videos coupled with the expedited delivery models of social media platforms gives deception, propaganda, and disinformation an entirely new meaning. Hence, raising awareness about DeepFakes and how to accurately flag them has become imperative. However, given differences in human cognition and perception, this is not straightforward. In this paper, we perform an investigative user study and also analyze existing AI detection algorithms from the literature to demystify the unknowns that are at play behind the scenes when detecting DeepFakes. Based on our findings, we design a customized training program to improve detection and evaluate on a treatment group of low-literate population, which is most vulnerable to DeepFakes. Our results suggest that, while DeepFakes are becoming imperceptible, contextualized education and training can help raise awareness and improve detection.

CCS CONCEPTS

• **Security and privacy** → **Human and societal aspects of security and privacy**; • **Computing methodologies** → *Machine learning algorithms.*

KEYWORDS

DeepFakes Detection, Human Perception, Detection Algorithms, Awareness

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI '21, May 8–13, 2021, Yokohama, Japan
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445699

ACM Reference Format:
Rashid Tahir, Brishna Batool, Hira Jamshed, Mahnoor Jameel, Mubashir Anwar, Faizan Ahmed, Muhammad Adeel Zaffar, and Muhammad Fareed Zaffar. 2021. Seeing is Believing: Exploring Perceptual Differences in DeepFake Videos. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3411764.3445699

1 INTRODUCTION

In the current "era of misinformation", AI has a pivotal role. DeepFakes in particular, which are videos that have been manipulated to alter their contents, are one of the major drivers of the spread and permeation of this misinformation. Generated by swapping the face of an actual person in a video with that of someone else (or by superimposing and merging the features as well as expressions of another person), these videos are becoming increasingly more realistic to the point that it is hard to distinguish them from authentic unadulterated videos. On the flip side, AI-based detection mechanisms are still playing catch-up [48]. Though progress is being made [13, 15, 22, 49], the technology behind the generation of DeepFakes, such as autoencoders [20] and Generative Adversarial Networks (GANs) [50], is more advanced. This implies that for the foreseeable future, DeepFakes will continue to have an impact on the fake narratives pushed forth on content distribution platforms, such as social networking applications and news outlets.

A sobering example of this impact of fake content was witnessed in India in April 2018, when a carefully doctored video of a child's abduction went viral on WhatsApp [3]. The video created national hysteria and the fallout was so severe that 9 people ended up dying in various tragic incidents across the country. Among those who were killed was a 55-year old woman who was lynched by a mob for handing out sweets to some children. In fact, the video created so much panic and agitation in certain underprivileged strata of the society that two men, who pulled over to ask for directions, were mistaken for the kidnappers and were beaten to death by a mob. It was later discovered that the original video was from a child-safety

awareness campaign from the neighboring Pakistan. However, as large sections of the society were unable to identify the video as doctored (particularly those who fell in the lower end of the digital literacy spectrum), precious lives were lost.

Similarly, the government of Gabon allegedly released a DeepFake video of President Ali Bongo on 1 January 2019, to quell rumors that he had passed away after the president had not been sighted for a few months [47]. The critics of the government immediately questioned the authenticity of the video and a week after the video was released, the military attempted an unsuccessful coup. The political turmoil witnessed by Gabon since then is a testament to how DeepFakes can impact the course of events at the national and international level. Similar to the upheaval caused in Gabon, numerous reports of DeepFakes being used to interfere in sensitive national affairs, such as influencing the US presidential elections [34], swaying political views in India ahead of Legislative Assembly elections [46], etc., have come to the limelight.

Going forward, the problem of DeepFakes only stands to grow further in severity. For instance, in a report that underscores the far-reaching impact of DeepFakes on our society, it was argued that the divisive technology will lead to an increase in child suicide rates in the future [27]. As more aspects of our existence and our identity transition to the cyber realm, we start to believe our self-created online personas and avatars. For teenagers, when these digital replicas are scandalized, such as through the non-consensual dissemination of a person's *DeepNudes*, or tarnishing of their reputation by circulating fake videos of an embarrassing rumor about them (loosing bladder control in front of their class for instance), they experience a deep shock. Depression and isolation follow, ultimately causing people to lose hope and take their own lives [5, 31]. The aforementioned examples serve as a stark reminder of how fake videos play at the very core of the deep-rooted sensitivities of our society, such as our love for our children, our conservative/religious values, and our political affiliations.

Recently, issues surrounding DeepFakes have started receiving substantial traction in the developed world. For instance, the US government passed the National Defense Authorization Act (NDAA) in 2019 [9], which has numerous provisions against election interference via DeepFake propaganda from nation state actors. As part of the law, the Director of National Intelligence (DNI) is required to submit a report to Congress anytime US intelligence agencies identify a potential national security threat arising from DeepFakes (or as mentioned in the law "machine-manipulated media" and "machine-generated text") [18]. In similar spirit, as part of its effort to combat the rise of disinformation, Facebook recently launched a "DeepFake Detection Challenge" with a prize money of $1 million at stake. The winner was announced in June of 2020 and reported an accuracy of around 65% [16] against a black box dataset of 100,000 videos. It is evident, that DeepFakes have rattled some feathers and that governments, along with tech firms, are coming together to combat against the threat the videos pose.

Unfortunately, discussion about DeepFakes is missing from the public discourse in some developing countries. Though individual efforts are being taken, a collaborative international undertaking has been largely missing. To prevent tragic incidents of India or Gabon from happening again, the public needs to be educated through targeted awareness campaigns against DeepFakes. However, in order for such training programs to be successful, we first need to answer two questions. First, what exactly should a person look for when determining the authenticity of a video. Or to put it differently, what makes a video fake and how to find this "fakeness". Second, is it possible to train people to increase the likelihood of detecting/identifying this "fakeness" in a video. Our work attempts to answer these questions through a combination of user-focused studies (surveys), eye gaze tracking technology and "unrolling" of deep learning-based detection algorithms. This "mixed-methods" approach allows us to determine what parts of a video frame users focus on when they consume media and how this makes detection of DeepFakes difficult or easier. On the other hand, the AI-based detection techniques enable us to "see" the regions that a machine focuses on when it attempts to detect if a video is fake or authentic. If we merge the findings extracted from the two datasets, some patterns emerge that can be leveraged to train people and make detection easier.

With the primary goal of creating awareness around the spread and permeation of DeepFakes, we make the following contributions in this work:

- **User study (Survey) dataset creation and analysis:** In our baseline study, we recorded data from a diverse population (N=95) regarding their DeepFake detection capability on videos generated from three different types of DeepFake generation algorithms. Moreover, we noted down their responses on encountering fake videos and documented their previous exposure to DeepFakes. We found that the detection accuracy was less than 26% on high quality DeepFakes. In addition to the accuracies, this dataset also contains user responses on regions that they focus on while attempting to detect if a video is fake or not.
- **Eye gaze dataset creation and analysis:** To compliment the user study where respondents were asked to highlight regions of interest in a video, we also created an eye gaze tracing dataset using CloudGaze [17]. During the baseline survey, when we asked the participants to classify a video as fake or real, the CloudGaze library ran in the background and captured the coordinates where the user was focusing on by analyzing the movement of their eyes. The resulting dataset is a rich collection of frame coordinates containing regions of relevance given a particular type of video. Together, the findings from the user study and the eye gaze dataset allowed us to develop a better understanding of human perception, particularly in regards to DeepFake detection and enabled us to fine-tune our training in the later stage.
- **Comparative study of human and machine perception:** As machines currently outperform human subjects in detecting DeepFakes (65% accuracy was reported in the DeepFake Detection Challenge organized by Facebook [16]), we wanted to develop a better understanding of why this is the case and if humans can learn something from machines. Hence, we implemented some popular deep learning-based algorithms from the literature and ran them on the same videos that were being shown to the participants of the user study. By using Class Activation Maps (CAMs or simply heat maps),

we were able to identify the regions that played a significant role in the detection of the DeepFakes and how these were different for the human test subjects.
- **Design and development of training course:** Based on the findings from the user study, the eye gaze tracking dataset and the comparative analysis between machines and humans, we designed a short training program (with a duration of 10 minutes) to raise awareness in our society about how to detect DeepFakes. We tested the efficacy of the training material on a low-literate sample (46 human subjects) from the population as this segment of the society is more likely to fall for a DeepFake. We found that the 23 subjects who received the training (treatment group) performed significantly better **(an increased accuracy of 33%)** than the other 23 subjects that did not (control group).

2 BACKGROUND AND RELATED WORK

We begin by providing an extensive discussion of the current landscape on DeepFakes. DeepFakes are synthetic images or videos created using deep learning techniques that look indistinguishable from real media [41]. While they have been around in some form since at least 1997 [6], recent advancements in deep learning have led to the creation of extremely high quality DeepFake generators [11, 26, 30, 35, 38, 50]. These state of the art generators produce synthetic videos that are so realistic and convincing that they can easily fool human viewers [38]. In most of these techniques, a *content generator network* is placed against a *discriminator network* that looks at the content generated by the generator and informs it of any visual "errors" it is making. When this goes on for thousands of iterations, the generator eventually learns to fool the discriminator and creates refined high-quality synthetic content that is visually hard to differentiate from real content [19].

While initial uses were limited to the adult entertainment industry, in the last few years, DeepFakes have been exploited for significantly more malevolent purposes. For instance, DeepFakes have been used to spread fake news [47], sway public opinion before elections by creating fake videos depicting candidates [40], and serve countless other malicious purposes. To make matters worse, a recent study on the diffusion patterns of viral media suggested that images and videos are likely to spread at a faster rate than news [12]. Clearly, the harmful potential of DeepFake technology and its aggressive diffusion patterns argue for well-orchestrated awareness campaigns to educate everyday media consumers about the dangers it poses. Indeed, in the age of information – and misinformation – DeepFake technology can be a powerful and dangerous weapon [1, 21, 32, 33, 43, 44].

In the next few paragraphs, we provide more detail on the methods for both DeepFake generation and detection. Finally, we will highlight a few cases where improving human perception by training people and raising awareness helps in combating DeepFakes.

2.1 DeepFake Generation Algorithms

Perhaps the earliest work on Machine Learning-based fake videos was published in 1997 [6], in which existing footage was manipulated to change the way a person uttered words to match a new script. Since then, DeepFakes have come a long way and the quality of videos produced by newer techniques is extraordinary, easily fooling humans [38]. These techniques include transformer-based frameworks, CNNs, and advanced face extraction, as well as specialized techniques to "polish" the final product for a seamless finish [39, 42, 51]. The most harmful DeepFakes enable the algorithms to "assume" a person's visual identity and make them do actions or say things as desired. *Reenactment* is used to replace facial features or landmarks, such as expressions, mouth, gaze, etc, or the posture and pose of a target with those of the source, while keeping the identity of the target intact [28]. Some techniques target individual facial features or poses while others compound the features and/or poses of the source to manipulate the target. *Replacement* is used to replace the face of the target with that of the source, either completely or driven by the target's features, maintaining the identity of the source all the while [28]. Numerous tools exist today that allow users to generate DeepFakes with little prior knowledge. A popular open source tool, *faceswap*, uses deep learning and provides a GUI for generating such DeepFakes [10]. Similarly, DeepFaceLab provides a simple framework with an easy-to-use pipeline for generating fake videos [36].

2.2 DeepFake Detection Algorithms

There has been a plethora of work in developing algorithms to detect DeepFakes. Most of these algorithms use different techniques in machine learning to build classifiers for detection. Popular techniques use facial features and landmarks [24, 38, 49], video inconsistencies [15, 25], image processing [22], and even video metadata [13] to identify DeepFakes. We categorize the approaches mentioned in the literature into the following bins:

Feature-Based Approaches: Some techniques target general issues in DeepFake generation algorithms to detect manipulated videos. For instance, Li et al. [25] used the rate of eye blinking in videos to detect DeepFakes. The key insight being that most DeepFake generation algorithms use training examples that contain pictures with open eyes, causing them to struggle in creating realistic eye blinking in the generated videos. Clearly, if everyday media consumers are made aware of this shortcoming, they can leverage it to flag a video as suspicious.

Methods based on Video Inconsistencies: Other techniques use inconsistencies between the face and the surroundings [15, 25] as features for DeepFake detection. Differences in camera views, lighting conditions, and video codecs of the base and target videos can sometimes lead to such inconsistencies. Again, if educated, some of these inconsistencies can be detected if a consumer pays close attention to details. On a slightly different pane, certain approaches [13, 15] rely on the lack of temporal awareness in the frame-by-frame generation process that can be detected by pixel-level CNN feature extractors.

Deep Learning-Based Approaches: More recently, researchers have started leveraging various types of deep learning methods, such as capsule networks [29], to detect DeepFakes. Popular approaches in this genre exhibit better accuracy however, due to their innate complicated architectures, the models are more challenging to interpret. At a high level, the approaches explore the difference in the orientation of copied and target faces [49], analyze traces left by transformations on the synthesized faces (e.g., scaling, rotation,

Figure 1: One of the ways in which DeepFakes are generated: source and target faces are gathered and a deep learning model is used to swap the faces such that the swapped image is indistinguishable from a real image.

shearing) [25], contrast the differences in the blood volume of the micro-vascular tissue of the cheeks and other central regions of the face [8], etc. Similarly, there have been studies that have leveraged the power of deep learning models but have approached the problem of DeepFake detection from a slightly different perspective. For instance, Koopman et al. [22] performed Photo Response Non-Uniformity (PRNU) analysis on video frames to create a fingerprint of the images. The PRNU pattern of a digital image is a unique noise pattern created by small factory defects in the light sensitive sensors of a digital camera and can be used to distinguish forged images from real ones. For the purposes of this work, we limit the set of deep learning techniques to those simple approaches that can provide value to the objective of educating users and raising awareness.

2.3 Discussion

Humans are increasingly having a hard time detecting DeepFakes [37]. Several large scale studies on the detection of DeepFakes have shown detection accuracy to be close to 50%, which is equivalent to random chance [45] [14]. While automated detection tools can prove to be useful, humans continue to be the weak link. Additionally, the most vulnerable strata of the society, who are characterized by limited resources and low digital literacy, often lack access to these automated tools, leaving a vast majority of people at risk. Most of these factors make DeepFakes a challenging problem to solve, and as mentioned before, one that can have dire consequences.

2.4 Motivation

Given the rate at which DeepFake technology is advancing, automated detection tools will soon face serious challenges. Of course, detection technologies are also evolving but shortcomings remain as *generation* is a step ahead of *detection*. Hence, the main goal of this study was to raise awareness around DeepFakes and investigate if users can be educated to be on the guard when consuming media. Along with automated detection approaches, this would provide an additional line of defense where users can use their own cognitive abilities to be cautious of suspicious videos. In addition to our primary goal of raising awareness around DeepFakes, we had three additional goals for this study; 1) to conduct an analysis of automated DeepFake detection tools and develop an understanding of how the detection is happening, 2) to conduct a user study to measure DeepFake detection on a representative sample from the developing world and 3) to use the results of step 1 to educate people, especially the low-literate, to be more mindful and vigilant about DeepFakes in their routine content consumption.

Figure 2: Visual representation of where algorithms and humans focus on when attempting to detect DeepFakes. a) Real Image b) Heatmap from the 'Ensemble of CNNs technique' [4] c) Heatmap from self-reported fake-looking features in the baseline survey d) Eye gaze data from our baseline survey

3 RESEARCH GAPS

Based on our study of the DeepFake problem domain, we have identified the following research gaps that require attention:

- **Perceptual differences in Humans vs Machine:** Although plenty of work has been done on individually analyzing the performance of humans and algorithms for DeepFake detection, no meaningful progress has been made to understand the perceptual differences between how algorithms and humans differ in their approach and detection methodology.
- **Effect of training on the DeepFake detection performance of humans:** The term DeepFakes describes videos that are generated using disruptive deep learning techniques. However, as they are targeted at a human audience, they also have a psychosocial aspect to them, especially when we talk about detection by humans and how to train the higher cognitive functions to be more vigilant of them. In our survey of the problem space, we have not been able to find any work that has explored the possibility of training people in a contextualized way to flag suspicious content and quantify the improvement in their ability to accurately detect DeepFakes.
- **DeepFakes and the developing world**: A substantial portion of the population in the developing world has limited exposure to the latest advancements in the field of AI and deep learning. As a result, they are not fully aware of how the technology is utilized to alter imagery. Hence, they can fall prey to the most simple video manipulations and fully believe that the video is authentic. The incidents in India and Gabon are testament to this (as both are developing countries) and lead us to conclude that there is a

serious lack of research in the developing world focusing on the DeepFake discourse. Hence, more effort is needed to investigate how people in the developing world view DeepFakes and whether they can be made more vigilant through awareness and training sessions.

Based on the identified gaps, we first explore and analyze the existing literature on DeepFake *detection*, coupling that with our user studies in the developing world, to identify the perceptual differences between humans and machines. Second, based on the findings from the surveys and the detection analysis, we develop a targeted training program aimed at raising awareness around DeepFakes and educating people on how to detect them. We observe significant improvement in the ability of our treatment group to correctly identify DeepFakes and show the results in later sections.

3.1 Value & Contribution to Community

Our work has two key objectives: a) create awareness on the existence of DeepFakes and b) investigate whether the most vulnerable among us can be educated to be more vigilant of this phenomenon. As stated earlier, the technology behind DeepFakes is constantly evolving and improving. However, the awareness regarding DeepFakes and how to recognize them is not increasing at the same rate. In fact, considering the ability to instantly share content across social media platforms, where virality of content can be taken as a proxy for veracity, we believe a comprehensive effort on part of all stakeholders (social media giants, content distribution platforms, news outlets, educational institutions, etc.) is required to educate people and create awareness around this phenomenon. As we will show in later sections, human subjects who severely lacked the ability to recognize DeepFakes became more sensitized following their participation in our training program and were "on the lookout" for fake content. Hence, our study underscores the fact that creating awareness is an effective strategy in battling against DeepFakes and will continue to remain relevant in years to come, even as more complex ways are devised to create fake content. In fact, we sincerely believe that the humble effort behind this work needs to be taken up and spun into a larger awareness campaign around misinformation and fake content. Given the devastating impact of DeepFakes, the value of the paper can be seen as a small step in the DeepFake discourse, which might potentially save a life or help prevent mass hysteria from a viral disinformation DeepFake.

4 EXPLORING AUTOMATED DETECTION TECHNIQUES

As mentioned earlier, several techniques exist for the automated detection of DeepFake videos. Some of these techniques rely on visual cues or indicators in the imagery to make the determination on its authenticity. For the purposes of this study, these cues are of paramount importance as we want to identify these indicators and perform a differential analysis between humans and machines. [1]

[1] As a passing remark, we recognize that the science behind DeepFakes is continuously evolving and the quality is improving with each new approach. Hence, purely visual cues-based detection might become more challenging in the future. However, as we have mentioned before, the consequences of a viral DeepFake can be so devastating that we feel that humans need to be made aware now, and educated on how to play their part in conjunction with the automated machine-based approaches (which could yet take some time to mature and fully eliminate this problem).

To this end, we selected four main automated DeepFake detection schemes to understand what parts or aspects of a video they focus on. We used these particular detection schemes because 1) their code was openly available and fully functioning, 2) they showed high-accuracy on standardized datasets (see Table 1 for the accuracies reported on such datasets), and 3) their results were easily interpretable. Again, the goal was not to encompass all state of the art deep learning approaches into our study (that would be a herculean undertaking), rather we just wanted to investigate if there is something humans can learn from deep learning techniques or not and hence, we went with a small representative sample of techniques popular in the deep learning community. We ran these detection schemes on our datasets and analyzed the results, which are summarized as Content Activation Maps (CAMs or simply heatmaps) in Figure 2. As depicted, the selected machine learning algorithms are treating various facial regions like eyes, mouth and nose of the person with more significance (highlighted by the red regions in the heatmaps) and hence, could potentially add value to how humans approach this problem. Below, we provide more details on the various techniques that were explored:

- *Exposing DeepFake Videos By Detecting Face Warping Artifacts [25] (Face Warping)*: The technique presented in this study uses the distinct artifacts in DeepFakes that are inevitably created as a result of resolution inconsistency between the warped face area and the surrounding context. The method detects such artifacts of significance by comparing the generated facial landmarks and their surrounding regions with a dedicated Convolutional Neural Network (CNN) model.
- *Video Face Manipulation Detection Through Ensemble of CNNs [4]*: The authors of this study use ensembles of CNNs for face manipulation detection. Their scheme relies on obtaining different models, starting from the base network, EfficientNetB4, using an attention mechanism and a Siamese training strategy. The second column in Figure 2 shows the effectiveness of the attention mechanism in extracting the most informative content from fake faces. The results demonstrate that the attention network again focuses on facial regions like eyes, lips, ears, nose, chin, etc.
- *Exposing Deep Fakes Using Inconsistent Head Poses [49]*: This technique (*Head Poses*) is based on the intuition that when a DeepFake is created by merging two faces, the position and alignment of the two faces can differ ever so slightly. This can lead to inconsistencies when 3D head poses are estimated from the face images. The authors use the dlib library to extract 68 landmarks on the face from the input image/video to estimate head pose. Then, a subset of these landmarks is again used to estimate head pose for the central region. Inconsistencies using the full set of facial landmarks and those in the central region are used to train the SVM classifier to differentiate DeepFakes from real images or videos.
- *FaceForensics++: Learning to Detect Manipulated Facial Images [38] FF++*: This technique processes an input image by a face tracking method, which is used to isolate the facial area. The result is fed into a learned classification network that outputs the prediction using an XceptionNet architecture [7].

Table 1: The reported performance on standardized datasets of the four detection techniques we analyzed. The values indicate that the techniques form part of those representative in the space of DeepFake detection.

Study	Classifier/Method	Best Performance	Dataset
Bonettini et al. [4]	XceptionNet	92.7% (AUC)	FaceForensics++, DFDC
	EfficientNet and EfficientNet with attention	94.4% (AUC)	FaceForensics++, DFDC
Rossler et al. [38]	XceptionNet Cropped Image	70.10% (Accuracy)	Face2Face, FaceSwap, DeepFakes, NeuralTextures
	Bayar and Stamm [2]	61.6% (Accuracy)	Face2Face, FaceSwap, DeepFakes, NeuralTextures
Yang et al. [49]	SVM at frame level	89.0% (AUROC)	UADFV, DARPA MediFor GAN Image/Video Challenge (subset)
	SVM at video level	97.4% (AUROC)	UADFV, DARPA MediFor GAN Image/Video Challenge (subset)
Li et al. [25]	ResNet50	97.4% (AUC)	UADFV, DeepFakeTIMIT
	ResNet101	95.4% (AUC)	UADFV, DeepFakeTIMIT

Table 2: Performance of respondents at identifying real vs. fake videos in the baseline survey. Precision, recall, and F1 scores are reported over the complete dataset. These scores cannot be calculated per dataset, since each dataset contains only positive or only negative examples.

Dataset	Accuracy	Precision	Recall	Fscore
DeepFaceLab [35]	25%	As noted, each dataset had a single class. Therefore, these metrics were only calculated on the overall results		
FF-DF [38]	58%			
Celeb-DF [26]	21%			
Real	88%			
All fake	**38%**			
Overall	**51%**	**81%**	**38%**	**54%**

5 BASELINE SURVEY

We carried out two user studies in this work. In this section, we provide the details of the first user study that we conducted. The goals of this study were to 1) raise awareness around DeepFakes among the most vulnerable strata of the society, 2) quantify the DeepFake detection capabilities of a representative population sample and 3) investigate the features of a given video that users instinctively focus on when trying to determine if it is fake or not.

5.1 Participant Recruitment and Survey Conduction Procedure

The study was conducted during the peak of the Covid-19 lockdown. Hence, participant recruitment was particularly challenging due to severe restrictions on movement, cumbersome social distancing policies, and strict sterilization and disinfection procedures. Therefore, we tried several different methods of recruitment to build a decent sized participant pool and avoid any potential recruitment biases. We primarily advertised our study through university emails and student-focused Facebook groups. Furthermore, we managed to recruit a number of undergraduate student volunteers to conduct the survey at home with their family members and other cohabitants. The students were given a set of clearly defined instructions about guiding the respondents through the survey and recording their answers - a step by step guide of the process, including relevant links to the website and questionnaire. A standard set of 'dos and don'ts' of conducting the survey were also shared as part of the instructions to make sure conductors follow best practices of conducting such user studies. These included statements such as: conduct the survey with one person at a time, do not reveal the nature of the survey to the respondent beforehand (to avoid priming), phrase instructions and questions in a manner that do not turn into leading questions, etc. Responses were recorded on paper and reported through our online web-based survey application.

Since we could not conduct a large scale survey physically, we made an online application for orchestrating the survey. The design of the application was kept minimal and intuitive with a user-friendly and simple interface to remove unnecessary distractions. We also embedded an eye-tracking software in our survey application to track which part of the image participants are focusing on when asked to distinguish between a real and a fake video. This is because the perception of inauthenticity in visual data can often be difficult to express or even consciously recognized in terms of discrete features. For this, we used GazeCloud [17], an open source eye tracking library, which uses webcam to track users' gaze. Gazes

were recorded in the background without any visual feedback to avoid any distractions when watching the videos.

5.2 Dataset of DeepFakes and Real Videos

Before explaining the step-by-step procedure of how the survey was conducted, we first describe the video dataset that was used in our user studies. We curated a set of 25 fake videos and 4 real videos from publicly available sources. The sources included FaceForensics++ [37], Celeb-DF [26], and DeepFaceLab [36]. Each of these sources used a neural network-based approach in generating the videos in their datasets. We briefly describe each one below:

- Celeb-DF [26] dataset contains high-quality DeepFake videos of celebrities with very few visual inconsistencies or artifacts. The generation algorithm uses a series of refinements in the synthesis process to reduce color mismatch, temporal flickering, and inaccuracy of extracted face masks, resulting in a polished finish with convincing imagery.
- FaceForensics++ (FF-DF) [37] dataset consists of a set of videos generated using four different automated manipulation techniques. We selected the videos created using the FaceSwap [51] technique, henceforth referred to as FF-DF. Faceswap uses two autoencoders with a shared encoder that are trained to reconstruct training images of the source and target. A face detector is used to crop and align images. To create a fake image, the trained encoder and decoder of the source face are applied to the target face. The autoencoder output is then blended with the rest of the image using Poisson image editing.
- DeepFaceLab [36] is the leading open-source face-swapping framework with an easy-to-use pipeline for generating fake videos. Briefly, the technique relies on an encoder-decoder pair to extract features from face A and connect it with the decoder of face B to reconstruct face B from the original face. Like the Celeb-DF approach, the dataset gathered here consists of high quality imagery, which appears to be very realistic and convincing.

We picked these sources based on the quality of the videos produced and their general popularity. We ensured that all the DeepFake videos were created by implanting an external face onto a subject in the original video. Real (unadulterated) videos were also taken from the same sources so as to match the themes in the fake videos. For instance, if we selected a DeepFake with a news reporter, then we also added a real video of a similar duration featuring a news reporter.

5.3 Methodology

We describe the detailed steps that were followed during the baseline survey below:

(1) The respondents were shown a set of four short videos, randomly picked from the pool of our dataset so that one of the videos was real and the rest were fake. In order to truly capture their instincts, the participants were not detailed about the task at hand. Rather, they were simply told to watch the videos, which would then be followed by a set of questions. We also provided the option to skip the video if the respondent had seen the content before. This was done to remove any bias due to prior knowledge about the video.

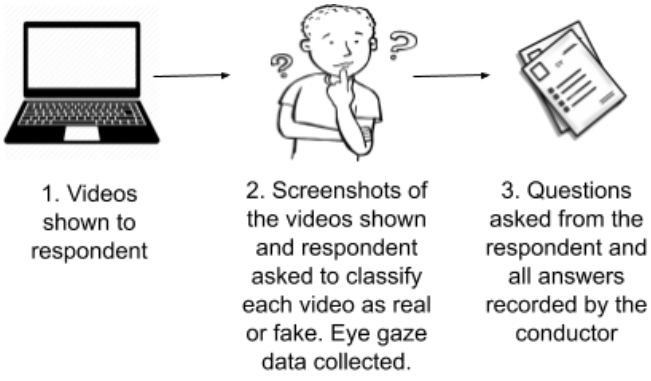

Figure 3: Stepwise description of the baseline survey

(2) After the respondents had concluded watching all the videos, they were then explicitly asked to differentiate the fake videos from the real videos (similar to what the automated classifiers do). To this end, we showed them four screenshots of the video (joined together in the form of a 2x2 grid) and left them on their screen till they made their decision. Furthermore, at this point, as the respondents focused on the screenshots to analyze the imagery and give their final verdict, we collected the eye gaze data in the background to find the points they were focusing on. The details of the eye gaze dataset are presented later in the section 5.6.
(3) The final component of the survey consisted of questions, which were asked by the conductors:
   (a) Based on the categorization of a particular video, the conductor was instructed to follow up with a pre-defined set of questions to determine the process that allowed the respondents to arrive at their conclusion. These included which features of the video seemed fake to the respondents and how they differed from their understanding of reality. While the conductors had a list of possible features on their questionnaires, they were kept hidden from the respondents to reduce priming.
   (b) A few general questions were asked about their demographic details, their internet usage and computer literacy, their level of concern about the spread of fake videos, as well as their knowledge of and previous encounters with DeepFakes. The conductors were also provided a space to note any interesting observations they made during the survey.

To avoid distractions and loss of attention, the questionnaire was kept short and concise. The language used was simple, and any hard jargon or unfamiliar terms for the respondents were avoided. The questions were to the point and clearly worded. Furthermore, the entire questionnaire was extensively reviewed by experts in user studies and questionnaire design to remove any leading or double-barreled questions that could result in biased or vague responses.

Each survey took about 10 minutes to complete and all surveys were completed in the period between August 1-31, 2020.

5.4 Ethical Considerations

Before we began working with human subjects, we obtained IRB approval for our survey methodology from our local institutional review board. We started each survey by informing the users of the type of data we will be collecting and that it will be anonymized. No video was recorded and only gaze data was collected. We then asked the user for their consent to participate in the anonymous data collection. We concluded each survey by educating the participants about DeepFakes and again obtaining explicit approval for the use of their survey responses.

5.5 Details of Participants and Results

There were a total of 95 respondents in the age range of 11-73, spread over 11 cities. 56% were male, 42% female, and 2% preferred not to disclose their gender. The student volunteers who were assisting with the user studies made sure that there is no duplication of a respondent.

Table 2 summarizes the results for our baseline survey. While our participants were able to recognize 88% of the real videos, they performed quite poorly on the fake videos. The detection accuracy for the different datasets varied from 21%-58% with low precision and recall. As discussed earlier, the Celeb-DF, and DeepFaceLab dataset consisted of higher quality videos with fewer color mismatches, temporal flickerings and a higher resolution. In fact, the results for the baseline survey are also inline with our assumptions that higher quality DeepFakes are difficult to detect for the general audience. Figure 4 shows the accuracy of the participants at detecting DeepFakes in our survey, on each of the videos in each of the dataset mentioned in section 5.2.

Surprisingly, our survey results showed no correlation between detection capabilities and age, gender, or education. We asked our participants about their level of comfort with the internet (on a scale of 1-5) as a proxy for digital literacy and also found no correlation between this self-reported number and detection ability. While 51% responded that they had seen DeepFakes before, we found no significant correlation between their responses and the results. Interestingly enough, male respondents were statistically more likely to claim they had detected DeepFakes in the past ($\phi = -0.301**, p < 0.01$). We had also asked the respondents how concerned they were about the spread of fake content, on a scale of one (not concerned) to five (very concerned). The average score was four. Furthermore, we had asked the conductors to note down interesting observations made during the surveys. A recurring theme was the complete inability to understand that a video could be fake beyond the content (fake news). Another point worth noticing was that respondents would confidently point out that movie scenes could not be fake (as some of the fake videos were derived from movie scenes).

5.5.1 Feature based results. Typical approaches for DeepFake detection rely on visual cues to separate fake videos from real. During our survey, we asked our participants to explain what factors led them to classify each video. The word salad of responses from our participants is summarized in Figure 5. Detailed questions to each participant allowed us to see what percentage of participants were looking at what visual cues for video classification. Two of the authors manually encoded the qualitative responses, classifying each response according to the feature(s) being talked about. We further analyze the results in Table 3. While there were differences between different datasets, most of the respondents focused on the background, eyes, forehead, lips, cheeks, and expressions when looking for visual cues. DeepfaceLab and Celeb-DF datasets were difficult to detect correctly and the results show that background and hair were important factors for classification. For lower-quality fakes produced by FF-DF, people were focusing on eyes, nose, and other facial features for inconsistencies.

5.6 Complimentary Dataset: Eye Gaze Tracking

In order to correctly identify the regions a user focuses on (in addition to their self-reported responses from the questionnaire), we used a javascript library (GazeCloud [17]), which tracked eye gazing of every participant and stored gaze fixations as coordinates on the screen at a frequency of 10 data points per second. We made Class Activation Maps (CAMs) or heatmaps from a utility provided by the same library in order to visualize the most frequented regions and ensure a uniform comparison with that of the analysis done on DeepFake detection tools earlier as well as the feature-based responses. The heatmap-based visualization coupled together the actual gaze data with user explanations to highlight the regions that users focus on while trying to detect DeepFakes. This allowed us to gain a deeper understanding of their cognitive classification process. The data acquired here was also used in making the training strategy, as described later in section 6.1.

5.6.1 Gaze Data Processing and Cleaning. To remove noise in the eye gaze data, we only used data collected in the first thirty seconds of a user seeing the four screenshots of the video. This is because we noticed that initially the users actually focus on the 4 screenshots for each video (the 2x2 screenshot grid), but, in most cases, they soon start looking at the conductor or their gaze moves away from the screen, as their higher-level cognitive processes start wrapping up and they start inclining towards one of the two categories (real or fake). We combined the data points from each of the four screenshots that were shown in the decision phase by aggregating them on a single image. We also removed the data points at the corners of the images to avoid using data points that could have appeared due to participants transitioning their attention to a different screenshot of the same video. Moreover, we used a minimum focus time threshold of half a second in a 50px area, which approximately translates to the size of a face in the screenshots. Finally, we aggregated the heatmaps from different participants who saw the same image. Around 60% of data points were filtered through this process.

5.7 Discussion on Baseline Survey Study

Our results from the first survey provide interesting insights into how humans perceive DeepFakes. As we will show in the subsequent sections, the data gathered in the first survey is instrumental in how we developed the training program and helped educate people in detecting DeepFakes. More importantly though, the survey helped raise awareness among the participants who were all

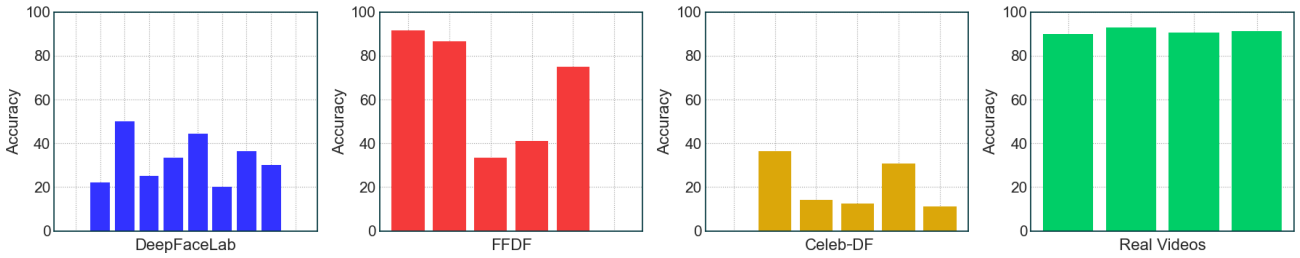

Figure 4: Accuracy of identifying DeepFake videos correctly, categorized according to the aforementioned generation techniques. Accuracy of correctly identifying the real videos is also shown. The videos of the DeepFaceLab and Celeb-DF datasets were relatively harder to identify as DeepFakes.

Table 3: Features that help identify DeepFakes along with percentages of respondents who identified them as classification aids in the baseline survey. These features were asked only when a participant classified a video as fake.

	DeepFaceLab [35] (N=23/96 correctly identified)	FF-DF [38] (N=71/122 correctly identified)	Celeb-DF [26] (N=14/68 correctly identified)	Real (N=10/95 incorrectly identified)	Combined Fake Identification % (N=108)
Hair	13.04%	4.23%	**21.43%**	9.09%	8.33%
Forehead	17.39%	**28.17%**	7.14%	0.00%	23.15%
Eyes	8.70%	**42.25%**	14.29%	18.18%	31.48%
Nose	4.35%	**25.35%**	14.29%	0.00%	19.44%
Lip	8.70%	**39.44%**	14.29%	9.09%	29.63%
Cheek	17.39%	**47.89%**	0.00%	0.00%	35.19%
Ear	4.35%	4.23%	7.14%	9.09%	4.63%
Chin	13.04%	11.27%	7.14%	9.09%	11.11%
Eyebrows	0.00%	**38.03%**	7.14%	0.00%	25.93%
Background	**39.13%**	**22.54%**	**42.86%**	18.18%	28.70%
Face	17.39%	14.08%	7.14%	9.09%	13.89%
Movement	13.04%	5.63%	7.14%	0.00%	7.41%
Body Language	8.70%	1.41%	7.14%	18.18%	3.70%
Skin Tone	0.00%	4.23%	0.00%	9.09%	2.78%
Lighting	8.70%	2.82%	0.00%	9.09%	3.70%
Context	17.39%	1.41%	**21.43%**	18.18%	7.41%
Expressions	13.04%	5.63%	**21.43%**	18.18%	9.26%
Clothes	4.35%	1.41%	0.00%	0.00%	1.85%

Figure 5: A word map showing the words used most commonly by respondents in the baseline survey. These have been extracted from free text answers that described the features respondents felt helped them distinguish between real and fake videos.

intrigued by the study and became more cynical, at the very least, towards media received through social media platforms and other content distribution frameworks. For instance, when shown a fake video, most respondents did not question its authenticity. In fact, most of our conductors were left in shock when numerous respondents failed to flag even poor quality DeepFakes let alone better ones. However, the mere participation in this exercise gave them enough exposure to DeepFakes that they promised to be mindful of the existence of the phenomenon going forward.

As a passing remark, we acknowledge that detecting DeepFakes through visual cues is an arms race. As soon as a feature becomes significant for detection purposes, newer models are trained to improve the quality and remove any "fakeness" indicators. Initial DeepFakes had clear inconsistencies with lip-syncing, skin-tone, facial expressions, etc., characteristics that could be leveraged for

detection. However, newer models are producing results that are far better in quality. Skin tone and eyebrows are examples where our participants were unable to detect any inconsistencies in higher quality videos as their synthesization has been improved and look more realistic. However, as our study highlights, raising awareness still helps in most cases and can prevent DeepFakes from gaining enough momentum to cause serious damage.

5.7.1 Differences in the Perception of Humans and Algorithms. With our analysis of automated (or Machine Learning-based) detection algorithms (described in section 4) and results from our baseline survey, we observed some differences in the way machines and humans perceive DeepFakes. As can be seen in Figure 2, there is a noticeable difference between the regions of the face that humans focus on as compared to those on which the DeepFake detection mechanisms focus. Humans were found to have focused on the major regions of the face, for example, eyes, hair, nose, etc., and typically overlook the minor details. On the other hand, the detection algorithms seemed to have also focused on fine-grained details, such as the location and structure of the ears, face outline, etc, perhaps because these automated mechanisms have the advantage of looking at the imagery at the level of individual pixels (and small clusters of pixels). We were also able to identify common mistakes that users make and the most useful strategies that people were using in classifying a given video. In summary, our results showed subtle but noticeable differences between humans and machine perceptions. In fact, the purpose of studying detection algorithms was to see if insights could be leveraged to help human learning. We believe that humans can be made aware of these differences to help them deal with DeepFakes and improve their overall approach to detecting DeepFakes. We used some of these differences to create contextualized training modules for our controlled experiments that are discussed in section 6.

6 SPECIALIZED TRAINING SURVEY

At a high-level, the goal for this study was to quantify the improvement in a person's ability to detect DeepFakes as a result of increased awareness through targeted training. To achieve this goal, we conducted a controlled experiment on a group of participants along with a training lesson, which is described in section 6.1. We then compared the responses of the group that underwent training (treatment group) to the group that did not (control group) on how accurately they are able to flag DeepFakes. Similar to the baseline study, IRB approval for this study was taken beforehand from our local institutional review board and participants were informed regarding the data collected, with their consent taken before and after the study.

6.1 Training Strategy

Since there is no prior work that details the best practices on training people to detect DeepFakes, we constructed our own custom approach by combining results from our baseline survey and the analysis of automated DeepFake detection algorithms (section 4). This approach allowed us to combine the best of both machine and human approaches to create effective awareness and training material. We created a set of clear walkthrough examples, derived mostly from the findings from the baseline survey and the automated approaches.

Part of the training involved walkthrough examples of both easy to detect DeepFakes and ones on the harder side of the detection spectrum (classification based on the results from the baseline survey). This was done to give our subjects some context on how DeepFakes have evolved in terms of quality. In addition to the DeepFakes, a real video was also included in each training session as a baseline to compare against and highlight the most common telling elements (significant features) that tend to stay consistent throughout. Using this mix, a case was made to highlight the importance of analyzing content rationally and objectively - especially in cases when the content is polarizing in nature. Additionally, each DeepFake example had detailed instructions with certain highlighted features and points of interest along with a corresponding analysis strategy to help identify the inconsistencies in the fake video. A detailed description of each example has been provided in Table 4. The features, as mentioned previously, were borrowed from the results of the baseline survey and insights from the detection algorithms. To be precise, we combined results of the detection algorithms (described in section 4) with the self-reported features (answers to open-ended questions) and eye gaze data from the baseline survey (described in section 5).

Intuitively, it can be argued that some of the features will of course be common across both humans and machines, and this was observed in our findings as well. For instance, 31.48% of the respondents reported focusing on 'eyes' as one of the more significant features in making a judgement. Similarly, we saw in the heatmaps that the automated detection approaches were also treating eyes as an important factor. However, other features, such as the 'chin' were only reported by roughly 11% of the participants. However, the chin or the jawline is actually a prominent feature that is examined by some of the automated approaches. In the training, we included features from both the user study and the DeepFake detection tools, as long as they were helpful and effective in raising awareness and improving vigilance against DeepFakes. Other important aspects such as 'blurriness', 'flickering' or 'skin tone' were adopted from the responses to the question 'How do the selected feature(s) compare to reality?' in the questionnaire section of the baseline survey. For instance, a participant who had selected 'eyebrows' and 'face' as the features, mentioned that "His [person in the video] face is glitching. Skin tone also varies near the forehead, above eyebrows". Another participant, having selected 'face' as the feature, mentioned "The face in the video was flickering, which does not happen in real life."

For convenience, the walkthrough examples were stitched together to form a slide deck, which served as a teaching aid during the training process. These slides included the major details that were to be communicated, starting from explaining what a DeepFake is (due to the lack of awareness regarding them) and moving on to a general explanation of what the person's first point of focus should be when trying to reason on the fakeness of a given video. These insights were obtained from the detection tools and the user study, as mentioned previously. For instance, some of the features that were explicitly pointed out to the treatment group included discoloration of the skin, random flickering on the face, certain blurry patches on various parts of the face, uneven or extra eyebrows, awkward lip movement, etc.

Table 4: Detailed description of each walk-through example. The first column represents the video example we used in the walk-through, the second column indicates signs of forgery in the video, and the third column indicates how we arrived at those signs (through data from the baseline survey or from the analysis of the detection tools we analyzed).

Video	Features to note and explanation provided	Adopted from
1	**Features**: Flickering eyes and face; Distortion on cheeks and lips. Such flickering and distortion are not common in genuine videos and so it is important to question the authenticity of such videos.	**Baseline Survey**: 35.19% on cheeks, 29.63% on lips, 31.48% on eyes **Tools**: eyes, lips
2	**Features**: Flickering on the face; Clear boundary on the forehead. It is important to note that the natural features be preserved in videos. The boundary on the forehead suggests that the person's face has been superimposed.	**Baseline Survey**: 13.89% on face, 23.15% on forehead **Tools**: boundaries of the face.
3	**Features**: Flickering of eyebrows; Blurriness on the eye region; Subtle color difference. Such differences in skin tone and blurriness (as compared to the rest of the face and background) should raise suspicion.	**Baseline Survey**: 31.48% on eyes, 29.63% on forehead, skin tone and lighting 6.48% **Tools**: eyes
4	No blurry regions; Eyes, eyebrows, forehead, jawline, cheeks etc stay consistent	Since the study and heatmaps pointed to eyes, eyebrows, cheeks, chin, and face outlines as the most commonly telling features, we decided to show how such features tend to be consistent in a real video
5	**Features**: Throughout the video, the chin seems to change - facial features are never changed and so must be looked at in cases of suspicion This is an example of a good DeepFake. While these are significantly harder to detect, the key takeaway here is to acknowledge their existence and be aware of the extent of potential deceptions. And be more cautious when approached with information, especially of polarized content.	**Baseline Survey**: 11.11% on chin **Tools**: chin
6	**Features**: Flickering and lighting differences on the face throughout, especially the eye region Notice how there is flickering and difference in lighting from time to time. Such aspects are not common and should not be ignored.	**Baseline Survey**: 13.89% on face, 31.48% on eyes **Tools**: eyes
7	**Features**: Blurriness on the central part of the face, especially the eye and eyebrow region as compared to the rest of the face Again, the blurriness in the central part of the face as compared to the rest of the face should raise suspicion.	**Baseline Survey**: 31.48% on eyes, 25.93% on eyebrows **Tools**: eyes
8	**Features**: Eyes and lips seem too blurry and unnatural as compared to the rest of the face When speaking, the lips of the person move unnaturally and become blurred. Similarly, one of the eyes is specifically blurry and seems unnatural as compared to the rest of the face.	**Baseline Survey**: 31.48% on eyes, 29.63% on lips **Tools**: eyes, lips
9	**Feature**: Eyes This is another example of a good DeepFake, with most of the features staying consistent. The only clue here is on the eyes in certain parts of the videos - while this is hard to detect, it is important to remain vigilant. Especially of content of a polarized nature.	**Baseline Survey**: 31.48% on eyes **Tools**: eyes

On the flip side, participants were also cautioned against relying on a few inconclusive factors that can lead to a flawed conclusion. For instance, a respondent had incorrectly identified a real video as fake based on the logo of a certain product in the videos, saying 'the

logo looks very unprofessional'. Some participants even based their argument on an pre-existing yet flawed understanding of a famous person's behavior: a participant claimed that 'hand gestures do not match those of [famous person in the video]'. Similarly, another mentioned how 'he did not place his hand on the chest'. Of course, such inconsistencies (fake logos, out-of-sync gestures, etc.), when present, can aid detection of DeepFakes. However, the training highlighted that the final judgement should not solely be based on such less relevant factors. Similarly, through a manual analysis of the eye gaze dataset, we found that participants who correctly identified DeepFakes had a tendency to focus more towards the face as opposed to the body of the person in the video. This phenomenon was noticed to be a repeating pattern, i.e., participants who focused more towards the facial region attained a higher accuracy of detecting fake videos. Thus, as part of the training takeaways, we asked the participants to focus more towards facial inconsistencies and, as secondary evidence, look for indicators in other regions (such as the body).

Of course, we do not claim that the training was comprehensive by any means (nor was that the objective of the study), rather the goal was to determine if users can be better prepared to analyze untrusted videos that they consume on a daily basis via content distribution and social media platforms. This study highlights the need for future work to address the potential of such an approach in helping spread awareness. Clearly, it would be of benefit to work on a more comprehensive study including a rigorous investigation of a larger portion of the representative state of the art generation and detection algorithms, and studying possible signals that could be of value in detecting DeepFakes. Similarly, more impactful training strategies can be developed in the future by leveraging a variety of explainable AI detection tools and extracting giveaway features from a larger sample of cutting edge DeepFake detection techniques.

6.2 Methodology

We conducted the survey on a sample of 46 participants (details in the next section), mostly comprising of low-literate population. As we wanted to target the most susceptible strata of the society, we went with a sample that was mostly located in the lower end of the literacy spectrum and had limited exposure to latest technologies. The following steps describe the user study and the training process:

(1) The respondents were divided into two groups: a control group (no training) and a treatment group (had training), dividing them such that each group had a similar occupation status.
(2) The **initial test** consisted of showing three videos (the pool consisting of videos from generation techniques, described in 5.2, and real videos) to both of the groups. This was done in a similar fashion to the baseline experiment, where they were not told the task beforehand so as to reduce priming.
(3) After the respondents finished watching the videos, they were shown four screenshots of each video (in a 2x2 grid) to help them recall. This was followed by a question to label them as fake or real. This step was done for both groups in the same manner.
(4) The treatment group was then given a thorough training on how to identify DeepFakes using the training material we had previously created (described in section 6.1). The conductor pointed out the relevant features in each video shown to the respondent, explaining why that is characteristic of a DeepFake. This was done to train the respondents to identify those features in DeepFake videos that are not present in an authentic video.
(5) The participants from both the treatment and the control groups were then given a **final test**, consisting of unseen fake and real videos for classification. All answers were recorded by the conductors as the videos were being shown. The ratio of fake and real videos was changed from the first step so as to make sure there was no bias from the previous round.

Note that due to the pandemic, the study was done individually with the participants, not in groups. Moreover, participants from the control group and the treatment group were requested not to discuss the study (methodology, steps, videos, etc.) with each other.

6.3 Participants and Results

In the second user study, there were a total of 46 respondents, 23 in the treatment group and 23 in the control group, in the age range of 18-55. 72% were male and 28% were female. We acknowledge that the sample contained more males, however this was a direct implication of the Covid-19 lockdown, and despite our best efforts we could not recruit more females. For the purposes of this study, the uneven division of males and females does not affect the major outcomes i.e., a) raising awareness helps in flagging DeepFake content and b) contextualized training can help combat their spread.

Table 5 summarizes the results from our controlled experiment. To begin, there was no correlation between the treatment and control groups in correctly detecting fake videos **prior to the training** ($r = 0.028$, $p = 0.854$). t-test of independent samples also revealed no statistically significant difference between treatment and control groups in detecting fake videos ($t = -0.185$, $p = 0.854$). Table 6 shows the mean, standard deviation, and standard mean error of the percentage of fake videos correctly detected by each group before training. There was also no statistically significant difference between groups based on age, gender, total score of detecting fake videos correctly or any of their scores on the videos independently. On the other hand, **after the training**, there was a statistically significant correlation between treatment and control groups in their ability to correctly detect fake videos ($r = 0.540$, $p < 0.001$). This means that participants in the treatment group were better able to correctly detect fake videos. Furthermore, independent samples t-test revealed a statistically significant difference ($t = -4.256$, $p = 0.000$). Table 7 shows the mean, standard deviation, and standard mean error of the percentage of fake videos correctly detected by each group after training. Before the training, the treatment and control groups were able to detect fake videos with an accuracy of 55% and 58% respectively, which is roughly similar. However, post training, the performance of the treatment group was significantly better compared to that of the control group (i.e., 88% vs 57% accuracy in identifying fake videos). More significantly, the treatment group was able to perform much better on the easier DeepFakes and the detection accuracy went from 57% to 80% for FF-DF. The control group's performance on the same set of videos

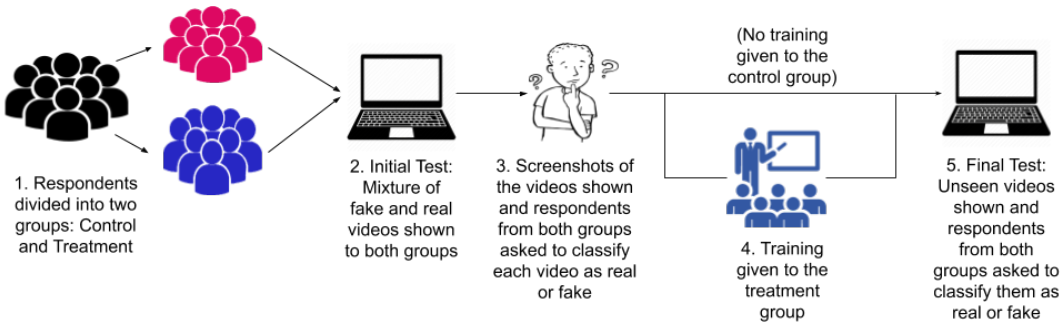

Figure 6: Flowchart describing the steps in the specialized training survey

Table 5: Respondents' accuracy for the lab study. There were a total of 23 respondents in each group.

	Control Group Initial Test	Treatment Group Initial Test	Control Group Final Test	Treatment Group Final Test	Absolute % Increase Control	Absolute % Increase Treatment
DeepFaceLab [35]	43%	39%	30%	35%	-13%	-4%
FF-DF [38]	57%	57%	54%	87%	-3%	+30%
Real	70%	74%	75%	65%	5%	-9%
All fake	**58%**	**55%**	**57%**	**88%**	**-1%**	**+33%**

Table 6: There was no statistically significant difference between treatment and control groups in detecting fake videos before training.

Group Statistics (Pre-treatement)					
	Group	N	Mean	Std. Dev	Std. Error Mean
Fake Detection Percentage	Control	23	0.359	0.190	0.040
	Treatment	23	0.370	0.207	0.043

actually went down from content 57% to 54%. It is also interesting to note that the treatment group was able to perform much better on the difficult DeepFakes as compared to the control group, and had a lower decrease than the control group. Note that the videos shown in the final test were different from the ones shown in the initial test, thus the relative difference between the control and the treatment group is meaningful rather than the absolute percentage increase/decrease between the two tests for each group. There was also a positive correlation between the treatment group and total Score (on all videos) of identifying fake videos (Cramer's Phi = 0.549, p = 0.031) - meaning participants were more likely to get higher detection score of fake videos in the treatment group.

6.4 Discussion on Training Study

Our goal for the training experiment was to create awareness among the individuals and prompt them to question the authenticity of videos that they come across in daily life. Furthermore, we also wanted to explore how training helps in this regard. The results show how the accuracy of the participants to identify fake videos increased significantly after the training, compared to the control group, where the accuracy remained somewhat same as before. Interestingly however, while both the groups had almost the same performance on real videos, the performance actually dropped by 9% for the treatment group post-training. This drop can be attributed to the fact that exposure to such fake content causes a lack of trust, resulting in a phenomenon called "Information Apocalypse" or "Reality Apathy" [23]. It shows how the participants were now more skeptical of all the videos they were shown during the experiment and were aware of the existence of such fake videos. This is important since now the participants can question the authenticity of videos and can at least identify the poor quality DeepFakes, raising the bar for the quality of DeepFakes used to spread misinformation.

7 GENERAL DISCUSSION AND LIMITATIONS

Role of Social Media and Messaging Apps: Once DeepFakes have been generated by malevolent actors for some nefarious purpose, the question about distribution still remains. Intuitively, it can be argued that content distribution platforms, such as Whatsapp, Telegram, YouTube and Facebook, due to their vast reach and unparalleled rate of spread are the perfect "unwilling" accomplices. Hence, it is often observed that viral fake videos are first shared on social media and messaging apps and from there make their way to more organized and federated content platforms. It is therefore, imperative to devise a crowdsourced strategy to recognize DeepFakes and stop them from spreading in their initial stages (as once a DeepFake has become viral there is very little that can be done).

Table 7: Participants in the treatment group were better able to correctly detect fake videos after training.

Group Statistics (Post-treatement)					
	Group	N	Mean	Std. Dev	Std. Error Mean
Fake Detection Percentage	Control	23	0.4804	0.2236	0.0466
	Treatment	23	0.7424	0.1927	0.0402

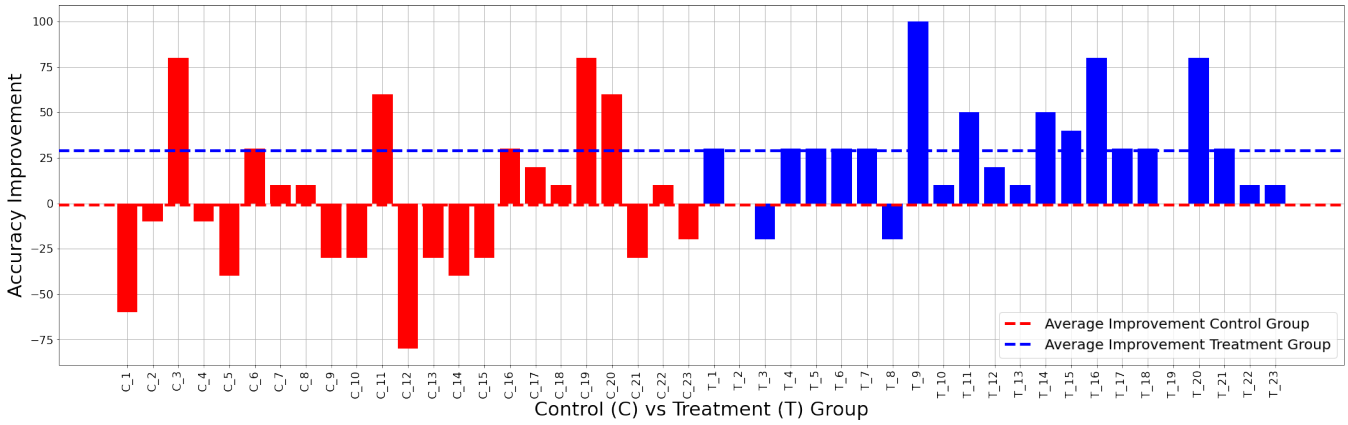

Figure 7: Improvement in identifying fake videos correctly, between the initial test and the final test. As shown, the treatment group experienced greater improvement.

Need for Training and Awareness: Like other aspects of computing where human beings are the weakest link, the spread and permeation of DeepFakes can be attributed primarily to a lack of understanding and misplaced trust on part of humans. Similar to the cybersecurity domain, technical solutions such as machine-assisted detection can help but eventually the onus falls on users to protect themselves. To this end, people need to be educated, trained, and equipped with the appropriate skillset to help mitigate the spread of DeepFakes. Currently, to the best of our knowledge, there are no educational and awareness initiatives that target issues surrounding DeepFakes. Hence, there is an urgent need for HCI researchers and practitioners to invest resources in this area and become more involved with the design of effective and interactive interventions like the one suggested in this paper.

Need for Technical Solutions: In addition to awareness campaigns, major tech companies like Facebook and YouTube should focus more on developing DeepFake detection algorithms so as to to filter videos from their platforms. Even though some initiatives have been taken, past incidents have shown that we are dangerously unprepared. In the absence of more effective filtration mechanisms, the fallout of a cleverly doctored viral video could be huge. With very little research being done on humans in this area, tech firms need to step up and compensate for this lack by investing heavily on their platforms to make sure they cannot be exploited. Again, our results from the comparative study can offer meaningful insights as to the most useful directions to pursue, in order to effectively mitigate this evil.

Limitation on Participation Recruitment: One of the primary limitations of our study was the selection of our participant pool. The work started when the pandemic was at its peak. As a result, a carefully crafted experimental strategy had to be abandoned. Instead, we tried to compensate by recruiting student volunteers to get to decent numbers. We tried to include participants from diverse backgrounds. However, again, the fact that a complete lockdown was in effect, made things challenging. Another limiting factor that affected our numbers was the length of the user studies and controlled experiments. Each session was around 15 minutes long (the 10 minute training was in addition to this), which meant that subjects were reluctant to participate given their busy schedules and conflicting engagements.

Selection of DeepFake Techniques: DeepFake creation and detection is evolving rapidly, and hence, the methods studied do not necessarily represent the cutting edge in these domains. However, our goal was not to analyze the state of the art generation or detection algorithms themselves. Rather, we wanted to leverage insights from automated algorithms to develop a training framework for raising awareness regarding DeepFakes. With this objective in mind, we chose the algorithms that had their code available online at the time of this study and could be analyzed under the hood to understand the cues they use for detection. Our work represents an effort to create awareness regarding DeepFakes, and teach people certain ways of determining whether a given video may have been manipulated. Clearly, a coordinated effort is needed in the future to include a larger and more comprehensive set of deep learning generation and detection techniques based on the state of the art. Similarly, the training also needs to follow from the findings of such a carefully orchestrated large scale study. The purpose of our work, which can be viewed as a pilot to a larger study, is to purely demonstrate that such an awareness-based approach can add value to the DeepFake discourse.

Eye Gaze Data Limitations: Another limitation came from the selection of Gazecloud [17] as the eye tracking software. Eye tracking does not currently work well on mobile devices, and the experiment was only performed on laptop and desktop computers with webcam support. Initially, the plan was to use specialized eye tracking hardware (in fact, the equipment was purchased and set up in a specialized lab environment), to track user perceptions of DeepFake videos. However, the lockdown forced us to abandon this route. For the future, we plan to conduct a full-scale study as soon as social restrictions are lifted using the eye tracking equipment.

8 CONCLUSION

In light of recent devastations caused by DeepFakes, it has become of paramount importance that society is educated about the dangers they pose and how to better identify them. Hence, in this work, we first conducted a user study of 95 respondents and asked them to identify DeepFakes as either real or fake and investigated what inconsistencies led them to their conclusion. While users were analyzing the videos from the survey, we tracked their eye gazes in parallel. This complementary dataset of coordinates, along with the regions highlighted by the users, allowed us to better understand which parts of a video users focus on and what are the cognitive processes that go behind the scene. In addition to this user study, we also implemented various deep learning detection algorithms to better understand the perceptual differences between the approaches taken by humans and those of machines. Based on the insights from the three datasets (user study data, eye gaze tracking data and the data from the comparative study of humans and AI), we designed a training program that walked users through various aspects of DeepFakes and how to better detect them. A controlled experiment was conducted to determine the efficacy of the training and if it had a meaningful impact on detection rates. Compared to the control group (no training), the detection capabilities of the treatment group (received training) increased across the board. We surmise that targeted awareness campaigns can help mitigate the risks associated with the spread and permeation of DeepFakes and can help safeguard our society against potentially devastating incidents.

REFERENCES

[1] C.R. Barnes and T. Barraclough. 2019. *Perception Inception: Preparing for Deepfakes and the Synthetic Media of Tomorrow.* Brainbox, New Zealand. https://books.google.com.pk/books?id=rCGnzQEACAAJ

[2] Belhassen Bayar and Matthew C. Stamm. 2016. A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer. In *A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer* (Vigo, Galicia, Spain) *(IH&MMSec '16)*. Association for Computing Machinery, New York, NY, USA, 5–10. https://doi.org/10.1145/2909827.2930786

[3] BBC. 2019. Deepfake videos could 'spark' violent social unrest. https://www.bbc.com/news/technology-48621452

[4] Nicolò Bonettini, Edoardo Daniele Cannas, Sara Mandelli, Luca Bondi, Paolo Bestagini, and Stefano Tubaro. 2020. Video Face Manipulation Detection Through Ensemble of CNNs. arXiv:2004.07676 https://arxiv.org/abs/2004.07676

[5] Marlee Bower. 2020. Loneliness and suicide: what's the link and what role does depression play? https://www.nationalelfservice.net/mental-health/suicide/loneliness-and-suicide-whats-the-link-and-what-role-does-depression-play/.

[6] Christoph Bregler, Michele Covell, and Malcolm Slaney. 1997. Video Rewrite: driving visual speech with audio. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1997, Los Angeles, CA, USA, August 3-8, 1997*, G. Scott Owen, Turner Whitted, and Barbara Mones-Hattal (Eds.). ACM, New York, NY, USA, 353–360. https://doi.org/10.1145/258734.258880

[7] François Chollet. 2017. Xception: Deep Learning with Depthwise Separable Convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Hawaii, 1800–1807. https://doi.org/10.1109/CVPR.2017.195

[8] Umur Aybars Ciftci and Ilke Demir. 2020. FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP (2020), 1–1. https://doi.org/10.1109/TPAMI.2020.3009287

[9] congress.gov. 2019. Deepfakes Report Act of 2019. https://www.congress.gov/bill/116th-congress/senate-bill/2065.

[10] deepfakes. 2017. deepfakes/faceswap: Deepfakes Sofware For All. https://github.com/deepfakes/faceswap.

[11] Apurva Gandhi and Shomik Jain. 2020. Adversarial perturbations fool deepfake detectors.

[12] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J Watts. 2016. The structural virality of online diffusion. *Management Science* 62, 1 (2016), 180–196.

[13] David Güera, Sriram Baireddy, Paolo Bestagini, Stefano Tubaro, and Edward J. Delp. 2019. We Need No Pixels: Video Manipulation Detection Using Stream Descriptors. arXiv:1906.08743 http://arxiv.org/abs/1906.08743

[14] Parul Gupta, Komal Chugh, Abhinav Dhall, and Ramanathan Subramanian. 2020. The eyes know it: FakeET–An Eye-tracking Database to Understand Deepfake Perception.

[15] D. Güera and E. J. Delp. 2018. Deepfake Video Detection Using Recurrent Neural Networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE Computer Society, Los Alamitos, CA, USA, 1–6.

[16] indianexpress.com. 2020. Facebook's $1 million 'Deepfake Detection Challenge' winner registers 65.18% accuracy. https://indianexpress.com/article/technology/tech-news-technology/facebook-deepfake-detection-challenge-winner-6460123/.

[17] Yoshio Ishiguro and Jun Rekimoto. 2012. GazeCloud: A Thumbnail Extraction Method Using Gaze Log Data for Video Life-Log. In *16th International Symposium on Wearable Computers, ISWC 2012, Newcastle, United Kingdom, June 18-22, 2012*. IEEE Computer Society, Newcastle, United Kingdom, 72–75. https://doi.org/10.1109/ISWC.2012.32

[18] jdsupra.com. 2019. First Federal Legislation on Deepfakes Signed Into Law. https://www.jdsupra.com/legalnews/first-federal-legislation-on-deepfakes-42346/.

[19] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Seattle, WA, USA, 8110–8119. https://doi.org/10.1109/CVPR42600.2020.00813

[20] Jan Kietzmann, Linda W. Lee, Ian P. McCarthy, and Tim C. Kietzmann. 2020. Deepfakes: Trick or treat? *Business Horizons* 63, 2 (2020), 135 – 146. https://doi.org/10.1016/j.bushor.2019.11.006 ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING.

[21] Jan Kietzmann, Linda W. Lee, Ian P. McCarthy, and Tim C. Kietzmann. 2020. Deepfakes: Trick or treat? *Business Horizons* 63, 2 (March 2020), 135–146. https://doi.org/10.1016/j.bushor.2019.11.006

[22] Marissa Koopman, Andrea Macarulla Rodriguez, and Zeno Geradts. 2018. Detection of Deepfake Video Manipulation.

[23] H. Li and Deepfake. 2019. The Emergence of Deepfake Technology: A Review.

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

[25] Yuezun Li and Siwei Lyu. 2019. Exposing DeepFake Videos By Detecting Face Warping Artifacts. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, June 16-20*. Computer Vision Foundation / IEEE, Long Beach, CA, USA, 46–52. http://openaccess.thecvf.com/content_CVPRW_2019/html/Media_Forensics/Li_Exposing_DeepFake_Videos_By_Detecting_Face_Warping_Artifacts_CVPRW_2019_paper.html

[26] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2020. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR June 13-19, 2020*. IEEE, Seattle, WA, USA, 3204–3213. https://doi.org/10.1109/CVPR42600.2020.00327

[27] Jim Meadows. 2019. Deepfake technology WILL lead to multiple child-suicides. https://medium.com/wearecommit/deepfake-technology-will-lead-to-multiple-child-suicides-d8a6535df992.

[28] Yisroel Mirsky and Wenke Lee. 2021. The Creation and Detection of Deepfakes: A Survey. *ACM Comput. Surv.* 54, 1, Article 7 (Jan. 2021), 41 pages. https://doi.org/10.1145/3425780

[29] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. 2019. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, IEEE, Brighton, United Kingdom, 2307–2311.

[30] Thanh Thi Nguyen, Cuong M. Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, and Saeid Nahavandi. 2020. Deep Learning for Deepfakes Creation and Detection: A Survey. arXiv:1909.11573 [cs.CV]

[31] Lu Niu, Cunxian Jia, Zhenyu Ma, Guojun Wang, Bin Sun, Dexing Zhang, and Liang Zhou. 2020. Loneliness, hopelessness and suicide in later life: a case–control psychological autopsy study in rural China. *Epidemiology and Psychiatric Sciences* 29 (2020), e119. https://doi.org/10.1017/S2045796020000335

[32] Carl Öhman. 2019. Introducing the pervert's dilemma: a contribution to the critique of Deepfake Pornography. *Ethics and Information Technology* 22, 2 (Nov. 2019), 133–140. https://doi.org/10.1007/s10676-019-09522-1

[33] Konstantin A. Pantserev. 2020. *The Malicious Use of AI-Based Deepfake Technology as the New Threat to Psychological Security and Political Stability.* Springer International Publishing, Cham, 37–55. https://doi.org/10.1007/978-3-030-35746-7_3

[34] Simon Parking. 2019. The rise of the deepfake and the threat to democracy. https://www.theguardian.com/technology/ng-interactive/2019/jun/22/the-rise-of-the-deepfake-and-the-threat-to-democracy.

[35] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr. Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, Sheng Zhang, Pingyu Wu, Bo Zhou, and Weiming Zhang. 2020. DeepFaceLab: A simple, flexible and extensible face swapping framework. arXiv:2005.05535 https://arxiv.org/abs/2005.05535

[36] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr. Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, Sheng Zhang, Pingyu Wu, Bo Zhou, and Weiming Zhang. 2020. DeepFaceLab: A simple, flexible and extensible face swapping framework. arXiv:2005.05535 [cs.CV]

[37] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2018. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179* abs/1803.09179 (2018), 21.

[38] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019.* IEEE, Seoul, Korea, 1–11. https://doi.org/10.1109/ICCV.2019.00009

[39] Yujun Shen, Bolei Zhou, Ping Luo, and Xiaoou Tang. 2018. Facefeat-gan: a two-stage approach for identity-preserving face synthesis.

[40] theconversation.com. 2020. Faked videos shore up false beliefs about Biden's mental health. https://theconversation.com/faked-videos-shore-up-false-beliefs-about-bidens-mental-health-145975.

[41] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. 2020. Deepfakes and beyond: A Survey of face manipulation and fake detection. *Information Fusion* 64 (2020), 131 – 148. https://doi.org/10.1016/j.inffus.2020.06.014

[42] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. 2018. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* IEEE Computer Society, Los Alamitos, CA, USA, 1526–1535.

[43] Cristian Vaccari and Andrew Chadwick. 2020. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society* 6, 1 (2020), 2056305120903408. https://doi.org/10.1177/2056305120903408 arXiv:https://doi.org/10.1177/2056305120903408

[44] Cristian Vaccari and Andrew Chadwick. 2020. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society* 6, 1 (Jan. 2020), 205630512090340. https://doi.org/10.1177/2056305120903408

[45] Cristian Vaccari and Andrew Chadwick. 2020. Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media+ Society* 6, 1 (2020), 2056305120903408.

[46] vice.com. 2020. We've Just Seen the First Use of Deepfakes in an Indian Election Campaign. https://www.vice.com/en_in/article/jgedjb/the-first-use-of-deepfakes-in-indian-election-by-bjp.

[47] washingtonpost.com. 2020. How misinformation helped spark an attempted coup in Gabon. https://www.washingtonpost.com/politics/2020/02/13/how-sick-president-suspect-video-helped-sparked-an-attempted-coup-gabon/.

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

[49] Xin Yang, Yuezun Li, and Siwei Lyu. 2019. Exposing Deep Fakes Using Inconsistent Head Poses. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019.* IEEE, Brighton, 8261–8265. https://doi.org/10.1109/ICASSP.2019.8683164

[50] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. 2019. Self-attention generative adversarial networks. In *International Conference on Machine Learning.* PMLR, PMLR, Long Beach, California, USA, 7354–7363.

[51] Jiangning Zhang, Xianfang Zeng, Yusu Pan, Yong Liu, Yu Ding, and Changjie Fan. 2019. Faceswapnet: Landmark guided many-to-many face reenactment.