SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY — INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Understanding The State of the Art of Publicly-Available Deepfake Detection Tools

Berdiguly Yaylymov

SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY — INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Understanding The State of the Art of Publicly-Available Deepfake Detection Tools

Der Stand der Technik bei der Erkennung von Deepfakes durch öffentlich zugängliche Tools

Author:	Berdiguly Yaylymov
Supervisor:	Prof. Dr. Jens Großklags
Advisor:	M.A. Severin Engelmann
Submission Date:	15.08.2023

I confirm that this bachelor's thesis is my own work and I have documented all sources and material used.

Munich, 15.08.2023 Berdiguly Yaylymov

Acknowledgments

First of all, I would like to thank my supervisor M.A. Severin Engelmann for his help and guidance throughout this thesis. Thank you for giving me the chance to write this thesis at the Chair of Cyber Trust and all your ideas and input. I am also grateful to Prof. Dr. Jens Großklags for giving me ideas and feedback during this project. Thank you Duc Trung Daniel Tran for proofreading and tips on the thesis. Finally, my sincerest thank you to all my friends and family who have stood by me and offered their support throughout not only this thesis but also my entire University journey.

Abstract

Deepfake technology, a fusion of deep learning and fake media, has rapidly evolved and become a powerful tool for generating highly realistic synthetic content. This advancement brings with it significant challenges in media authentication, entertainment industry, and privacy. As deepfakes become more sophisticated and accessible, the need for effective detection tools has become paramount. This thesis aims to provide a comprehensive understanding of the state of the art of publicly-available deepfake detection tools.

The study begins with a literature review that explores the evolution of deepfake technology, the various methods used for deepfake generation, and the existing approaches for deepfake detection.

A solid methodology is used to collect and study data on the existing tools. They are evaluated based on factors like precision, speed, accessibility, and ease of use. The selected deepfake detection tools are assessed in detail to provide insights into their features, capabilities, and performance.

The findings of this study highlights the pros and cons of the tested deepfake detection methods. By comparing them, we understand their unique features and how well they identify deepfakes in various media. The research also points out current issues in deepfake detection and suggests directions for upcoming studies.

This research has consequences across various areas such as media, entertainment, and legal matters. Recognizing the difference between real and manipulated content is vital for protecting the integrity of information, preserving trust, and fighting against false information. The knowledge shared in this research contribute to the ongoing efforts to develop effective deepfake detection mechanisms.

In conclusion, this thesis provides a comprehensive overview of publicly-available deepfake detection tools, offering a thorough evaluation and comparison of their features and capabilities. The study highlights the need for ongoing research and development in the field of deepfake detection to counter the growing threat posed by synthetic media. By promoting a deeper understanding of the state of the art in deepfake detection, this research aims to contribute to the advancement of techniques that can effectively mitigate the risks associated with deepfakes and synthetic media.

Contents

Acknowledgments **iv**

Abstract **v**

1 Introduction **1**
1.1 Background and Motivation 1
1.2 Thesis Structure 3
1.3 Objectives of the Study 4
1.4 Scope and Limitations 6
1.4.1 Scope of the Study 6
1.4.2 Limitations of the Study 6
1.4.3 Delimitations of the Study 7

2 Literature Review **8**
2.1 Techniques Used in Deepfakes 8
2.1.1 Autoencoders 8
2.1.2 Generative Adversarial Networks 9
2.1.3 Variational Autoencoders 10
2.2 Publicly Available Deepfake Generation Tools 12
2.2.1 DeepFaceLab 12
2.2.2 FaceSwap 13
2.2.3 Stable Diffuison 14
2.2.4 Neural Textures 15
2.2.5 FaceApp 16
2.3 Ethical and Legal Concerns 17
2.4 Existing Countermeasures and Detection Methods 18

3 Methodology **20**
3.1 Selection Criteria 20
3.2 Evaluation Metrics 21
3.3 Datasets 25
3.3.1 FaceForensics++ 25
3.3.2 Deepfake Detection Challenge Dataset 25

3.3.3 Face Forensics in the Wild . . . 26
3.3.4 OpenForensics . . . 26

4 Analysis of Publicly-Available Deepfake Detection Tools 28
4.1 Deepware . . . 29
4.2 Seferbekov . . . 31
4.3 NtechLab . . . 33
4.4 Facetorch . . . 34
4.5 Illuminarty . . . 36
4.6 AI or Not . . . 36

5 Case Studies 39
5.1 Entertainment and Art . . . 39
5.2 Politics and Media . . . 40

6 Results 41
6.1 Comparative Results of tools . . . 41
6.2 Final Results . . . 42

7 Discussion and Conclusion 46
7.1 Summary . . . 46
7.2 Future Work . . . 46
7.3 Conclusion . . . 47

Abbreviations 48

List of Figures 50

List of Tables 51

Bibliography 52

1 Introduction

The rapid and continuous development of Artificial Intelligence (AI) has given birth to numerous applications that have pushed the boundaries of what we previously believed to be possible. This thesis will delve into one of the most fascinating and alarming developments in this field, deepfakes. This work aims to help readers in understanding how deepfakes are identified, along with their current limitations, and potential future research directions.

1.1 Background and Motivation

In an era where digital media forms the cornerstone of communication, the advent of deepfakes, AI-enabled synthetic media, poses an unprecedented challenge to information integrity. Deepfakes, a portmanteau of 'deep learning' and 'fake' [39], [85], [93], is a technology that manipulates or fabricates audio-visual content to make it appear real, often indistinguishable from the original [83]. An example of a generated and altered image can be seen in Figure 1.1

Figure 1.1: Deepfake of Bill Hader impersonating Arnold Schwarzenegger. Screenshot from [20]

The proliferation of deepfake technology become initially sparked with the aid of its software in creating misleading movie star images and videos, before quickly expanding into different sectors. One of the earliest examples that drew widespread interest to deepfakes was a video made by an anonymous Reddit user named 'deepfakes' in late

2017 [38], [85]. This consumer started out to publish digitally altered pornographic motion pictures, realistically swapping the faces of actresses onto the bodies of porn stars. While these explicit videos were quickly removed, the sudden emergence of this facial replacement method, quickly caught the media's eye and circulated on various online forums [3]. However, it didn't take long for the technology to be applied beyond explicit content.

A notable example that showcased the potential of deepfakes, and arguably delivered it to mainstream attention, turned into a video of former U.S. President Barack Obama, released in April 2018 by Buzzfeed and Jordan Peele [13], [44]. The video features a deepfake of Obama announcing matters he by no means clearly stated, with Peele providing the voiceover. This deepfake video, effectively highlighted the potential misuse of this technology in spreading misinformation and propaganda.

In recent years, the sophistication of deepfake technology has reached an unprecedented level. An interesting example of this progression can be seen in the creation of 'Tom Cruise deepfakes' that circulated on social media in early 2021 [84]. The videos, created by Belgian visual effects artist Chris Ume in collaboration with actor Miles Fisher, who impersonated Cruise's voice and mannerisms, were shared on TikTok under the account name @deeptomcruise[1] [1], [143]. These deepfake videos show the synthetic 'Tom Cruise' doing various activities — performing a magic trick, playing golf, or simply telling a story about Mikhail Gorbachev [32].

The 'Tom Cruise deepfakes' took the internet by storm due to their uncanny resemblance to the real actor, in terms of both appearance and behavior. Unlike the early deepfake videos, which often exhibited glaring imperfections, these deepfakes were so convincing that many viewers initially believed they were watching the actual Tom Cruise. This level of realism underscored the strides made in deepfake technology, while simultaneously highlighting the potential dangers of its misuse.

Driven by advances in machine learning, especially deep learning, deepfake technology has grown significantly in sophistication and accessibility. The potential applications of deepfakes range from benign, such as in film production and entertainment, to malicious uses, including disinformation campaigns, identity theft, and deepfake pornography. As these applications become more widespread, deepfake technology has raised profound questions and challenges for society, especially regarding media authenticity, privacy, and cybersecurity.

However, it is not just the creation of deepfakes that has improved; strides have also been made in detection. There are now more sophisticated, AI-powered tools that can analyze videos and images for signs of manipulation. These tools operate on multiple levels, from detecting inconsistencies in lighting and shadows to looking for

[1] `https://www.tiktok.com/@deeptomcruise`

signs of digital artifacts and abnormal facial movements. But as detection tools become more sophisticated, so too do the techniques used to create deepfakes. This constantly evolving technological arms race underscores the critical need for ongoing research and development in deepfake detection.

In response to these challenges, there is an increasing need for robust and reliable deepfake detection tools. However, despite the flurry of research and development in this area, a comprehensive understanding and evaluation of the available detection tools remain elusive. This knowledge gap not only impedes the technological advancements in deepfake detection but also complicates the task of policy-making and regulation in this sphere.

This thesis is motivated by the need to bridge this gap and advance our understanding of publicly-available deepfake detection tools. By examining these tools, this study aims to contribute to the ongoing efforts to mitigate the risks associated with deepfakes and uphold the integrity of digital media.

1.2 Thesis Structure

Understanding the structure of this thesis is essential for a thorough understanding of the research, as it follows a logical and systematic progression. It starts by laying the basic groundwork, then gradually delves deeper into the specifics of the study, eventually integrating the findings and projecting forward-looking discussions. Below is a detailed outline of the thesis structure, which serves as a roadmap for navigating the document.

This initial chapter lays the foundation for the thesis. It provides an overview of deepfakes, introduces the topic of deepfake detection, and outlines the significance and timeliness of the study. It presents the objectives of the research by addressing three research questions, clearly stating what the study aims to achieve. The scope and limitations are also discussed here, delineating the boundaries of the research and acknowledging its constraints. The introduction serves as a guide, setting the reader's expectations for the rest of the thesis.

The literature review provides a survey of the existing body of knowledge related to deepfakes and their detection. The section begins with explanation of the techniques used to create deepfakes, giving the reader an understanding of the technology behind them. This section also highlights the ethical and legal concerns surrounding deepfakes and the countermeasures and detection methods currently in place. By identifying gaps and shortcomings in the existing literature, this section also underscores the relevance and value of the present study.

Chapter three, the methodology along with evaluation metrics and datasets adopted

for the study are outlined. The chapter provides information on how the publicly-available deepfake detection tools were selected for analysis. It also discusses the evaluation metrics used to test the effectiveness of these tools and the datasets used for testing. By detailing these elements, the chapter ensures that the research process is transparent and replicable.

Chapter four offers an analysis of the six deepfake detection tools: three for video and three for images. These tools were tested using metrics from Chapter 3, looking at their functionality, their advantages, and any drawbacks.

The fifth chapter takes the analysis from theory to practice, exploring real-world instances where deepfakes and their detection have played a significant role. The case studies are chosen to represent a variety of scenarios, thereby providing some of the practical implications and challenges associated with deepfakes and their detection.

The sixth chapter presents the findings from the evaluation of the selected tools. It also provides a comparative analysis of the tools, highlighting their relative strengths and weaknesses. Subsequently, the results of the comparison are described and discussed. The research questions from Section 1.3 are also answered here.

The final chapter summarizes the findings and discussions from previous chapters and reflects on their contribution to the field. It also identifies potential directions for future research, offering suggestions for how the field can continue to evolve and adapt in response to the dynamic nature of deepfakes. Finally, a conclusion is drawn, highlighting current deepfake limitations.

1.3 Objectives of the Study

Deepfakes have been prominently discussed in both scholarly articles and the media [122], [98]. Their ability to convincingly deceive an ordinary user is getting better and better. For effective countermeasures and understanding of future threats, it's important to understand the deepfake detection technologies. Thus, the primary purpose of the thesis is to provide an exploration of the state of the art in publicly-available deepfake detection tools. Specifally, the study addresses these Research Questions (RQ):

- Research Question 1: How accessible and user-friendly are publicly-available deepfake detection tools for individuals who do not have expertise in deepfakes or detectin deepfakes?
- Research Question 2: How effective are deefake detection tools, which are selected and analyzed in this research, in identifying forgeries from various deepfake generation tools?

- Research Question 3: What are the privacy implications and policies associated with selected deepfake detection tools?

To address the first research question, a methodological approach was adopted. The selection criteria, encompassing factors like accessibility, user-friendliness, and documentation are explained. These criteria guide the selection of the six tools examined in this study. Subsequently, the evaluation metrics are established to assess the strengths and shortcomings these detection tools. Initially, all tools undergo an evaluation based on the selection criteria outlined in Chapter 3. The assessment considers factors like the tool's accessibility, determining if it's open-source or proprietary. The availability of support and documentation, including installation guides and usage instructions, is also examined. The presence or absence of source code, difficulty of use, any internet or specific software and hardware requirements, and whether the tools are free or paid, are also considered. These evaluations are detailed further in Chapter 4.

In response to the second research question, the study evaluates the performance of the selected tools against a range of deepfakes generated through diverse methods. Initially, each of the six tools undergoes testing to discern if they can detect deepfakes. The video detection tools are evaluated using 110 videos, comprising 40 deepfake videos from DeepFaceLab [99], 40 from FaceSwap [70], 10 deepfake videos sourced from the FaceForensics++ [103] dataset, and 20 authentic videos from the same dataset. Image detection tools are assessed using a total of 123 images: 49 deepfakes produced with FaceApp [80], 54 with Stable Diffusion [115], and 20 genuine photographs obtained online. Futhermore, the training models and techniques utilized, the datasets employed for initial tool testing, and the programming languages and frameworks are also examined to answer this question.

Regarding the third research question, attention is directed towards the privacy measures of the selected tools, examining the presence of their privacy policies and their user data handling methods. This involves examining the potential risks and determining if the tool developers provide Terms of Use and Privacy Policy for their tools' utilization.

The identification and elaboration of these research questions provide a roadmap for the study, with each one serving as a crucial stepping-stone toward the understanding of the nature of the selected publicly-available detection tools. The research questions are then answered in Chapter 6, comparing strengths and limitations of the selected tools.

1.4 Scope and Limitations

The study of deepfakes and their detection is a broad field, involving a range of complex and interrelated topics. Therefore, it is essential to define the specific scope and limitations of this thesis to clarify what it will and will not cover. These boundaries not only provide clarity but also help ensure that the research is feasible and can delve into the chosen topics in sufficient depth.

1.4.1 Scope of the Study

The primary focus of this thesis is on the analysis and evaluation of publicly available deepfake detection tools. It will cover both the technical and societal aspects of these tools, including their performance, methodologies, implications, and potential areas for future development. It will also provide an overview of the current state of deepfake technology, from its historical development to its modern techniques and applications.

The thesis will mainly concentrate on visual deepfakes, encompassing both images and videos, aiming to offer an understanding of the deepfake environment. The research will also explore the dual nature of deepfakes, examining their harmless and harmful applications. This exploration is crucial to understand the difficulties involved in detecting deepfakes.

1.4.2 Limitations of the Study

Despite its broad scope, the study is subject to several limitations that should be acknowledged. Firstly, due to the rapid pace of technological advancements in the field of deep learning and AI, the state of the art in deepfake technology and detection tools can change swiftly. As a result, while the thesis aims to provide an up-to-date overview of the field, some of the information might become outdated shortly after publication.

Secondly, given the focus on publicly-available tools, this thesis might not capture the full spectrum of deepfake detection methodologies. Many sophisticated tools and techniques might be proprietary or classified information, not accessible for public use or scrutiny. Thus, while this study will provide an overview of the available tools, it might not cover the absolute cutting edge in deepfake detection.

Thirdly, while the study aims to objectively evaluate the performance of deepfake detection tools, it's important to note that this evaluation is based on the available datasets and metrics. Variations in these datasets, such as the quality and diversity of the deepfakes included, can impact the results. Moreover, no single evaluation metric or dataset can fully capture the effectiveness of a tool in all real-world scenarios.

Fourthly, while the study will explore the societal, ethical, and legal implications

of deepfakes and their detection, a full analysis of these complex and evolving issues is beyond its scope. These aspects will be discussed primarily in relation to the main focus of the thesis — deepfake detection tools — and may not cover all the potential implications of deepfakes.

Finally, the study is limited by the inherent challenges associated with deepfake detection. Deepfakes are a result of advanced AI and machine learning techniques, and detecting them is a complex task that is still an area of active research. Therefore, the study's findings should be viewed in light of these inherent difficulties.

1.4.3 Delimitations of the Study

While limitations are factors that are out of the researcher's control, delimitations are boundaries set by the researcher. In this study, due to time and resource constraints, the analysis will be limited to a few publicly-available deepfake detection tools, rather than a list of all available tools. Similarly, while the study will discuss a few illustrative examples of deepfake applications and case studies, it will not provide a comprehensive review of all possible uses or instances of deepfakes.

By acknowledging these scope, limitations, and delimitations, this thesis aims to provide a clear exploration of publicly-available deepfake detection tools while being transparent about its boundaries and potential areas of uncertainty.

2 Literature Review

The literature review sheds light on the understanding of deepfakes, their history, the technology driving them, the publicly available tools that create them, and the ethical, legal, and societal issues they raise. Additionally, it examines the countermeasures that have been developed to detect and deter them.

2.1 Techniques Used in Deepfakes

Deepfakes are underpinned by significant advancements in artificial intelligence (AI) and machine learning (ML), particularly the areas of deep learning and neural networks. Central to the creation of deepfakes are two techniques: autoencoders and Generative Adversarial Networks (GAN).

2.1.1 Autoencoders

Autoencoders, are a type of neural network used for learning efficient encodings or representations of input data [47]. In 2006, Hinton et al. [46], [47] raised the concept of deep learning. This idea quickly gained significant attention, making deep learning a primary focus of research worldwide [113]. The structure of autoencoders has two main parts: an encoder, which simplifies an image into a latent space[1] [65], and a decoder, which rebuilds the image using that simplified representation [65], [129]. Deepfakes use this setup with a common encoder that translates a person into this latent space. This simplified representation captures essential details about their face and body stance. This can subsequently be decoded using a model specifically trained for the intended target. So, in other words, during the encoding phase, it extracts facial details from two distinct individuals, identifying and storing the shared facial characteristics. Subsequently, in the decoding phase, it presents the information of these two individuals. The encoding of one person can be swapped with another, enabling the face of one person to be superimposed onto another in an eerily realistic manner.

The figure in Figure 2.1 presents an image going into an encoder. The outcome is a simplfied version of that face, ofeten called a latent face [152]. Depending on the setup,

[1] Latent space is a compressed representation of data where essential features are captured in lower dimensions [133]

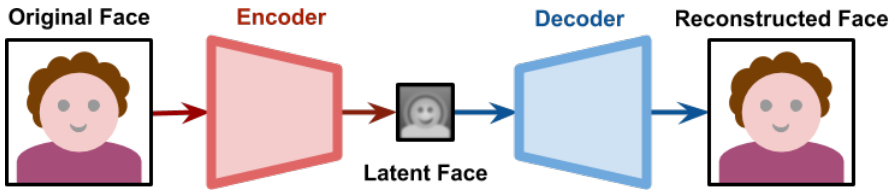


Figure 2.1: Autoencoders common structure. Figure from [152]

this latent face might not resemble a typical face. However, when it goes through a decoder, it's turned back into a face.

An example of using autoencoders is Fakeapp, developed by a Reddit user for deepfake generation [111], [45]. The process requires encoding and decoding pairs to interchange faces between input and output images. Distinct image datasets train each pair, but the encoder parameters remain consistent across both networks. Consequently, both encoder pairs utilize the same network. Given the consistent facial features such as the eyes, nose, and mouth across various images, this method allows the encoder to easily recognize similarities between two sets of facial images.

2.1.2 Generative Adversarial Networks

In the field of AI and Machine Learning (ML), two primary methods exist: supervised and unsupervised learning. The former relies on labeled data[2] for predictions, whereas the latter does not. Noteably, each method has its distinct advantages and nuances [26].

Supervised learning is a machine learning approach which uses labeled datasets to instruct algorithms in data categorization or predict outcomes. By using labeled inputs and outputs, the model calculates its precision and refines itself over time [7]. Supervised learning can be categorized into two data mining tasks: classification and regression. Classification involves accurately sorting data into specific groups, like distinguishing between newspapers and magazines. Meanwhile, regression uses algorithms to discern the relationship between dependent and independent variables [26].

Unsupervised learning applies algorithms to group and analyze unlabeled datasets, identifying hidden patterns without requiring human guidance [7].

Among the concepts and techniques used in ML, discriminative and generative models stand out as two widely used approaches [71]. So a *generative model* assesses the distribution of a dataset to determine the probability of a specific example. On the contrary, a *discriminative model* predicts unseen data using conditional probability and is applicable for both classification and regression tasks [42].

[2] In machine learning, data labeling involves attaching meaningful tags to raw data so that models can learn from it [5].

There are two main types of generative models, GANs and Variational Autoencoders (VAE) [71]. GANs introduced by Goodfellow et al. [41], envolve two competing neural networks: a generator and a discriminator. The generator produces fake samples and the discriminator distinguishes between the real and fake samples. Within a given training dataset, this method is adept at producing new data that mirrors the statistical properties of the training data. For instance, when taught using pictures, a GAN can create new images that look very real and similar to the original pictures [131]. While GANs were initially introduced as generative models for unsupervised learning, they have demonstrated utility in semi-supervised [106], fully supervised [56], and reinforcement learning [48] contexts.

At the heart of GANs lies the principle of 'indirect' training via the discriminator. This iterative process improves the quality of the generated samples over time as the generator learns to create more realistic fakes to fool the discriminator. This arms race pushes the boundaries of what GANs can create, contributing to the production of deepfakes that are increasingly difficult to detect [10].

GANs with alternative architectures. Various variants of GAN structures and associated generative models exits. In the original paper [41], GANs were designed using Multiplayer Perceptron Networks (MLP) and Convolutional Neural Networks (CNN). Later on, many other generative models and architectures, as described in Table 2.1, were introduced [9].

2.1.3 Variational Autoencoders

Recent developments have seen the rise of VAE and their use in deepfake generation [67]. Unlike traditional autoencoders, VAEs introduce a probabilistic spin to the encoding and decoding processes. The encoder, often designed using neural networks like the feedforward convolutional network, learns to encode input data into a latent space. The decoder, also based on a convolutional neural network, then reconstructs the original input from this latent space [90]. This allows for the generation of new faces by sampling from the learned distribution, enhancing the ability of the deepfake technology to generate entirely new, but convincing, faces.

While both VAEs and GANs are used for image generation, their methodologies differ. One primary distinction is their training approach. VAEs use an unsupervised learning technique, aiming to maximize the likelihood of the generated output relative to the input and compress the input into a latent space. Conversely, GANs are trained in a supervised learning technique, striving for equilibrium between the generator and discriminator, where the former seeks to fool the latter. Furthermore, VAEs often have a more straightforward training process compared to GANs because they don't require tight coordination between their components. Additionally, due to their advanced

Table 2.1: Some of GANs with alternate architectures [9], [131]

Types of GANs	**Description**
Conditional GAN (CGAN)	If both the generator and discriminator of GANs are conditioned on supplementary information, y, the model becomes conditional. This additional data, y, might be class labels or data from different sources. To condition the model, y is inputted into both the discriminator and generator as an extra input layer [86].
Dual GAN (DGAN)	A version of GAN, named DualGAN, uses two networks trained concurrently using two sets of unlabeled images. One network is designed for image generation, while the other distinguishes between generated and actual images. DualGAN effectively learns two image translators, making it suitable for diverse image-to-image translation activities [142].
Stack GAN (StackGAN)	A modified version of GAN uses multiple generators stacked together to create a more lifelike image. Stacked GANs compose a network designed to produce high-quality images [146].
Cycle GAN (CycleGAN)	CycleGAN is a method for automatically converting images from one domain to another, without the need for paired data samples [150].
Superresolution GAN (SRGAN)	A GAN designed to transform low-resolution images into high-resolution outputs. Super-resolution GANs use a combination of deep networks and adversarial networks to enhance the clarity of the input data [76].
Deep convolutional GAN (DCGAN)	A GAN that employs deep convolutional networks for both the generator and discriminator. This GAN relies solely on convolution and deconvolution layers. Studies suggest that images produced by the DCGAN architecture exhibit notably reduced noise [101].
Self-attention GAN (SAGAN)	The SAGAN facilitates attention-based, extended-range dependency modeling in image creation. In SAGAN, cues from all feature areas can be utilized to produce details. Additionally, the discriminator ensures that intricate details in separate parts of the image are consistent with each other [145].

capabilities, GANs are often used for demanding tasks such as super-resolution and image-to-image translation. On the other hand, VAEs are predominantly used for image denoising and generation [90].

2.2 Publicly Available Deepfake Generation Tools

As deepfake technology has evolved, so too has the ease of access to this technology. There are now several deepfake generation tools that are freely available and relatively easy to use, drastically lowering the bar for entry into the world of deepfakes.

2.2.1 DeepFaceLab

DeepFaceLab[3] [99], [59] was developed to address the challenges and inefficiencies typically observed in deepfake generation models [111]. This refined framework, facilitates face-swapping [99]. The architecture incorporates an 'encoder' and 'destination decoder', separated by an 'inter' layer, and concludes with an 'alignment' layer. To extract features, it uses the 2DFAN [12] heat map-based facial landmark algorithm, and for face segmentation, it utilizes the TernausNet [51].

Figure 2.2: Deepfake of Ivanka Trump impersonating Emma Watson. Screenshot from our own generated deepfake video dataset with DeepFaceLab.

Known for offering greater functionality and control over the deepfake creation process, DeepFaceLab has been used in several high-profile deepfake videos. Its sophisticated technology combines the power of GANs and autoencoders, leading to highly realistic face swaps in videos. DeepFaceLab offers tools for every step of the deepfake creation process, including face extraction, training, and video creation. This

[3] `https://github.com/iperov/DeepFaceLab`

comprehensive suite of tools, combined with its high-quality results, make it a popular choice among deepfake creators.

2.2.2 FaceSwap

Faceswap typically refers to the replacement of one person's face with another's in images or videos. In this context, it signifies a specific technique implemented in [70]. The method operates frame-by-frame for both source and target videos until one ends. Within each image, distinct facial landmarks such as face contour, eyes, mouth, and other distinguishing facial features are identified. Utilizing the landmarks from the source image, a 3D representation of the source actor's face is constructed. This is then adjusted to align with the target actor's facial landmarks and integrated into the target image. After processing each frame, the result is a video wherein the target actor's face is substituted by the source actor's [61], [60], [69].

Figure 2.3: Deepfake of Emma Stone impersonating Scarlett Johansson using Faceswap. Screenshot from [22].

This community-based deepfake tool stands out with its open-source nature, providing the user with a choice of multiple AI models. It caters to varying levels of experience and computing resources, making it accessible to a wide range of users. Besides its technical merits, FaceSwap[4] emphasizes the ethical use of deepfake technology, warning against non-consensual use of a person's likeness. It is more than just a tool; it's a community where people can learn, discuss, and share knowledge about deepfakes.

[4] `https://faceswap.dev`

2.2.3 Stable Diffuison

Recently, image synthesis methods using diffusion models centered on denoising techniques [49] have gained popularity because of their impressive outcomes in generating artificial artwork [72]. Notably, latent diffusion models like Stable Diffusion (SD) [102] empower individuals to produce images from textual prompts effectively, even using personal computers.

Introduced in 2022 [102], Stable Diffusion is a deep learning model utilizing diffusion methods, primarily for generating intricate images based on textual inputs. It also has applications in inpainting[5], outpainting, and facilitating image-to-image conversions guided by a text prompt [50]. The model was a collaborative effort from the CompVis Group at Ludwig Maximilian University of Munich [135], Runway, Stability AI[6], and several non-profit organizations [114], [88], [116].

As a latent diffusion model, Stable Diffusion is a type of deep generative neural network. Its code and model parameters are publicly available [18], and it's compatible with consumer-grade hardware having a GPU with a minimum of 8 GB [135]. This is in contrast to previous models like DALL-E [96] and Midjourney [55], which were restricted to cloud-based access [135], [124].

Figure 2.4: Deepfake of Justin Trudeau created with Stable Diffusion. Screenshot from our own generated deepfake video dataset with Stable Diffuison.

Stable Diffusion models utilize a diffusion process to generate realistic synthetic images from a simple Gaussian noise, achieving impressive results in the generation of deepfake images [138]. One of the benefits of diffusion models is that they can capture complex, multi-modal distributions in a way that other generative models may struggle

[5]Inpainting refers to the restoration technique where absent or damaged sections of an artwork are replenished to display a complete image [132].

[6]`https://stability.ai/stablediffusion`

with. This makes them particularly well-suited for tasks like deepfake generation, where capturing the detailed, multi-modal distribution of human faces is essential.

2.2.4 Neural Textures

Introduced by Thies et al. [119], Neural Textures represent a method for the storage, transmission, and rendering of learned neural representations in the context of computer graphics. By rendering with learned features instead of geometric detail, Neural Textures allow for more efficient representations, enabling high-quality, photorealistic image synthesis and editing. Specifically, in the realm of deepfakes, Neural Textures, as shown in Figure 2.5, can be trained to synthesize person-specific details, resulting in high-quality face swaps or manipulation of facial expressions in videos.

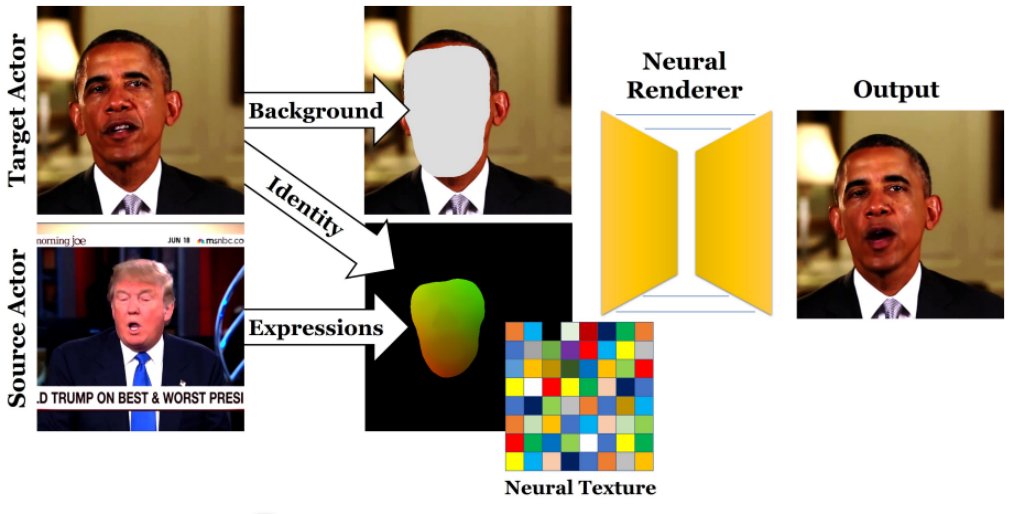


Figure 2.5: The reenactment synthesis process uses expression transfer to generate a UV map of the target actor that reflects the source actor's expression. This map, along with a background image, is processed by a neural renderer to create the final reenactment. Expression alteration is achieved by training a unique neural texture and renderer for the target actor, resulting in a manipulated video as shown in the image from [119].

Neural textures utilize a UV-map[7] connecting elements in the texture map to object points, allowing the creation of a viewpoint-specific texture. Integrating this into a trained deferred neural renderer produces an image from the designated viewpoint,

[7]UV mapping involves projecting a 3D model's surface onto a 2D plane for the purpose of texture mapping [136].

as illustrated in Figure 2.6. While neural textures have various uses, Thies et al. [119] highlight facial video manipulation. Figure 2.5 demonstrates this application, paralleling the Face2Face [120] method. Training videos establish a 3D face model of the target actor. Additionally, a unique neural texture and deferred neural renderer are crafted for this actor. Leveraging the Face2Face expression transfer technique, a UV-map from the target video is adjusted to mirror the source actor's expressions. With this UV-map, the neural texture, and renderer, the manipulated video emerges.

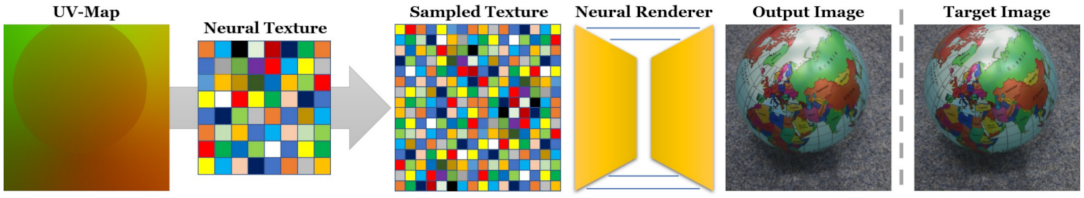


Figure 2.6: Overview of neural rendering pipeline [119].

2.2.5 FaceApp

While not a deepfake tool in the traditional sense, FaceApp[8] has gained popularity due to its ability to transform photos of faces in various ways, such as aging, de-aging, gender swapping, and adding smiles [137]. Introduced in 2017 [137] by the Russian startup Wireless Lab, now known as FaceApp Technology Limited, the FaceApp software enables users to undertake detailed photo and video modifications. These include aging or rejuvenating facial appearances, merging two faces, incorporating intricate facial expressions like smiles, and utilizing the debated 'gender swap' function.

The tool leverages neural network technology for its transformations, leading to surprisingly realistic results that have significantly contributed to the broader conversation around the manipulation of digital imagery [91]. FaceApp has been both praised for its technological achievements and criticized for its potential privacy and consent issues [126], reflecting the wider debates surrounding the ethical implications of deepfake technology.

In essence, FaceApp's image processing capabilities go beyond just simple pixel-altering image filters. Initially, each image is transformed into a multi-dimensional vector which might subsequently serve as a foundation for modification — essentially a reshuffling of the neural weights in the network [95]. When using FaceApp's tools, the CNN superimposes specific features onto the chosen portrait or selfie, features that have been extracted from its training dataset. Thanks to sophisticated image

[8] `https://www.faceapp.com/`

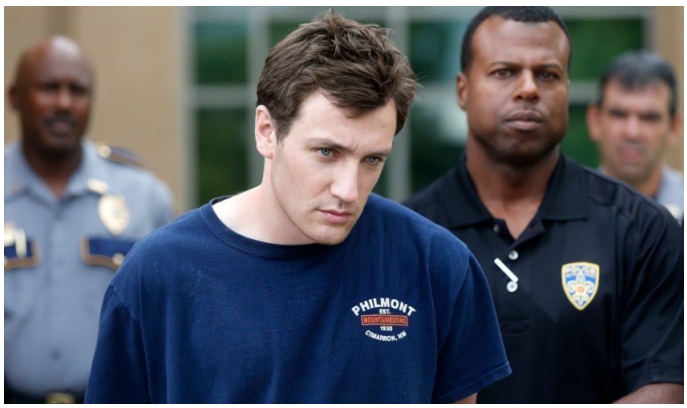


Figure 2.7: Deepfake created with FaceApp. Screenshot from our own generated deepfake image dataset with FaceApp.

recognition techniques, precise automatic feature adjustments are made, resulting in notably photorealistic effects. Intriguingly, while fundamentally transforming the image, FaceApp preserves unique facial features [14]. This allows users to perceive alterations, like aging or rejuvenation, as authentic modifications to their own faces [137], [100].

2.3 Ethical and Legal Concerns

The rise of deepfakes has brought with it a number of ethical and legal concerns. At the forefront is the issue of consent, as deepfakes often involve the use of a person's identity without their permission. This has been particularly common in the creation of deepfake pornography, leading to significant harm and distress for the individuals involved [15].

There are also concerns about the potential misuse of deepfakes in spreading disinformation and propaganda. Deepfakes could be used to manipulate public opinion, interfere in elections, or even incite violence [62], [77]. The realistic nature of deepfakes makes it difficult for the average viewer to distinguish truth from fake, further exacerbating these risks.

In journalism, the rise of deepfakes presents both a significant challenge and an ethical dilemma. Journalists must not only cope with the complex task of verifying the authenticity of deepfake content but also think about the ethical implications of using AI-generated content in their reporting. Misuse of deepfakes could lead to the spread of false information, shaking people's faith in the media [123].

Furthermore, the use of deepfake technology can be used to fabricate evidence in legal cases, potentially leading to miscarriages of justice. As deepfakes become

increasingly indistinguishable from real videos, the legal system will need to find ways to authenticate digital evidence and mitigate the risk of deepfake-generated evidence [15].

The business sector is not immune to the impact of deepfakes either. Businesses could fall victim to deepfake scams, in which AI-generated audio or video is used to impersonate a company employee or other authority figure. These scams could lead to significant financial losses or damage to a company's reputation [89].

Lastly, in deepfake detection, a crucial concern is the issue of false positives and negatives. A false positive, where a real video is wrongly flagged as a deepfake, could have serious consequences, such as the unnecessary spread of panic or unwarranted damage to an individual's reputation. On the other hand, a false negative, where a deepfake is not detected and is thus taken as genuine, can lead to the propagation of disinformation or fraud. These challenges underscore the need for highly careful deepfake detection methods [103].

2.4 Existing Countermeasures and Detection Methods

The rise of deepfakes demand effective countermeasures and detection methods to ensure information integrity and maintain public trust in digital content. As deepfake generation techniques have become more advanced, detection approaches have adapted accordingly. Many of these approaches use machine learning, particularly deep learning techniques, leveraging similar technologies that power deepfakes to combat them.

The general idea of deepfake detection is based on identifying inconsistencies or anomalies that typically arise during the process of creating deepfakes. These can be artifacts left by the specific algorithm used, unusual patterns in the distribution of the pixel values, or unnatural physical characteristics such as inconsistent lighting or improper blinking patterns [2].

One approach to detect deepfakes is frequency-based analysis, where the focus is on the differences in frequency patterns between original and deepfaked videos. These methods, like the one proposed by Durall et al. [30], take advantage of the fact that, deepfake generation algorithms usually function in spatial domain [105]. As a result, they may produce distinct inconsistencies within the frequency domain.

Another widely used approach is the CNN based detection. This type of deep learning model has shown excellent performance in various image and video processing tasks due to its ability to learn hierarchical patterns in the data. For deepfake detection, CNNs can be trained to learn the differences between real and fake images or videos, thus distinguishing deepfakes from the original media [92].

Another promising approach is the use of autoencoders for deepfake detection.

Autoencoders are a type of neural network that are trained to reconstruct their input data. By training an autoencoder on a large amount of real face data, it can learn to recreate real faces very well, but struggle to recreate deepfakes, allowing the detection of deepfakes based on the reconstruction error [19].

Recent advancements have led to the development of deepfake detection techniques that analyze physiological signals. For instance, Li et al. [79] developed a method based on the observation that real videos contain physiological signals that are driven by blood flow, such as heart rate. These signals, they found, are not well preserved in synthetically generated data and thus provide a new cue for deepfake detection.

It's important to note, however, that as deepfake generation techniques continue to evolve, the effectiveness of these detection methods can diminish. The constant race between deepfake creation and detection presents ongoing challenges for researchers and developers in maintaining the efficacy of these countermeasures.

3 Methodology

3.1 Selection Criteria

The core aim of this research study involves examining and evaluating various publicly available deepfake detection tools. It becomes imperative to establish a well-defined set of selection criteria for these tools. Selecting the right criteria is essential not just for representing a variety of detection methods, but also for making a fair comparison between the tools. The objective is to ensure that the evaluated tools are understandable, encompassing the nuances of accessibility, ease of use, limitations, variety in detection methods, and documentation. These are the criteria used to select the tools:

Ease of Use and Limitations. The tools were chosen based on their ease of operation. Considerations included the installation process of the chosen tool, the availability of Application Programming Interface (API), and whether a user with IT skills could utilize these APIs for integration purposes. Questions were raised about the specific operating systems needed, the requirement of a constant internet connection, and any inherent limitations of the tools. Do the developer or owner of the tool make any specific promises or claims about its use? Additionally, do the tools have restrictions on file soze or video length, and do these factors matter to use the tools?

Accessibility. Questions about the tools' accessibility include their public availability, any special account requirements for access, and if they can run on standard PCs or laptops. Can the tools be accessed via a website or is specialized software platforms like GitHub, Hugging Face or Google Colab or an Integrated Development Environment (IDE) required? The need for enhanced Graphics Processing Unit (GPU) or other hardware specifications, as well as the nature of the tool (open source vs. proprietary), were also considered.

Support and Documentation. Factors taken into account included the provision of user support, installation guidance, usage documentation, and resources for development. The existence of troubleshooting resources, community discussions for Q&A, and availability of the source code for developers' adaptability were important considerations.

Difficulty of Use. The complexity of the tools was evaluated on different levels. It was labeled as *Easy* if the tool was web-based with a user-friendly interface for detecting deepfakes. *Moderate* if it provided a platform like Hugging Face Space or APIs, and

Challenging if users had to devise their own way to utilize the tool.

Cost Considerations. Determined whether the tool is free or comes with a subscription fee, emphasizing a preference for freely available options. Also, if paid versions existed, did they provide additional functionality or ease of use?

Privacy Policy. The presence of privacy policies and terms of use for the tools was verified, underscoring their importance for tool legitimacy and trustworthiness. Especially when users are required to log in, it's essential to clarify how personal information is managed. Additionally, statements regarding ethical issues and adherence to relevant laws and regulations further bolster the tool's credibility and users' confidence in its operations.

The aforementioned criteria that were established played a critical role in the selection process of the tools. Tools were preferred that could be easily accessed and trusted by a broad audience. By ensuring tools had clear easy installations and user-friendly interfaces, the safety and simplicity of the selected tools were prioritized. It wasn't just about identifying tools that were publicly available; the goal was to ensure the chosen tools covered not just accessibility needs, but were also user-friendly enough for individuals without deepfake expertise. This approach aimed to make the tools both dependable and user-friendly, addressing a variety of user requirements.

An overview of the selection criteria is provided in Table 3.1.

3.2 Evaluation Metrics

After careful selection of the tools based on the prescribed criteria, it is important to systematically evaluate them to ensure their, accuracy and efficiency. This evaluation isn't just about how these tools perform; it's about understanding their strengths and potential weaknesses.

Each tool has a unique set of features and algorithms that drive its functionality. However, to compare them on a fair level and to ensure a comprehensive assessment, a standard set of evaluation metrics is applied. These are the evaluation metrics that were selected:

Processing Time and Scalability. The time taken for an input to be analyzed and authenticated by the tool is important. An average processing time, represented average seconds, was used in this assessment. The tool's ability to process multiple images or videos simultaneously and subsequently produce an output list indicating the authenticity of each piece of content was also examined.

Interpretability. The quality of the output produced by a tool was assessed based on its clarity and simplicity. Inputs were checked if they were simply labeled as 'Real' or 'Fake', or if a probabilistic estimate indicating the likelihood of manipulation was

Table 3.1: Selection Criteria

Selection Criteria	**Description**
Ease of use and Limitations	The tool's user-friendliness is determined by the simplicity of its installation process and its operational requirements. Is it a straightforward drag-and-drop mechanism, or does it demand any special software? Additionally, any constraints, such as file size limits or video duration caps, play a role in its overall user-friendliness.
Accessibility	Only publicly-available tools were considered to ensure broad user accessibility. The ease of use, from browser access to local installation, and hardware demands, with preference for standard configurations or cloud solutions like Google Colab, were pivotal in evaluations.
Support and Documentation	Proper documentation and active support, be it community or developer-driven, are crucial. Regular support ensures that users can fully make use of tool features, troubleshoot issues, and gain deeper insights into the tool's functionalities.
Difficulty of Use	The tool's complexity was categorized as *Easy* for web-based interfaces, *Moderate* for those using platforms like Hugging Face or offering APIs, and *Challenging* when users needed a custom approach.
Cost considerations	It is assessed if the tool is free or subscription-based, favoring free options, and checked if paid versions offered enhanced features or usability.
Privacy Policy	Tools were evaluated for privacy policies and user terms, emphasizing trust. It's vital to know how tools handle login data and their stance on ethics and legal compliance to ensure credibility.

provided. When probabilities were presented, the format (whether in percentages) and the threshold at which content was termed as a deepfake were investigated. In this study, content was classified as a deepfake if a probability of 50% or higher was observed.

Detection Accuracy. The fraction of true outcomes (including both true positives and true negatives) within the entire dataset was analyzed. This served to understand the tool's aptitude in correctly differentiating genuine from manipulated content.

Precision. The fraction of true positive outcomes within the total predicted positives (including both true and false positives) was evaluated. The tool's capacity to minimize false positives, ensuring that genuine content wasn't incorrectly flagged, was assessed.

Recall. The efficiency of the tool in correctly identifying genuine deepfakes from all the presented deepfakes was evaluated. The actual true positives that were successfully identified by the tool were calculated.

F1-Score. The F1-Score is the harmonic mean of precision and recall, providing a single metric that balances the trade-off between correctly identifying positive instances (precision) and the ability to detect all potential positive instances (recall), especially useful in situations where the class distributions are imbalanced[1].

Dataset Compatibility. The adaptability of a tool, based on the datasets it could align with, was explored. Tools that could function with a variety of datasets or those that came recommended with particular datasets were chosen.

Training Models. The types and quality of training models and techniques that were employed by each tool were explored.

Face Detection Techniques. The array of face detection techniques that were used by these tools was examined. Additionally, it was determined whether the tools used alternative detection methods.

Programming Language and Framework. The programming languages and frameworks that were important in the tool's development were reviewed to gain insights into its structure and potential functionalities.

Potential for Further Development. The possibility of whether tools could undergo extended development was investigated. Furthermore, the guidelines or requirements provided for such advancements, if any, were examined.

These metrics serve as a guiding light, distinguishing the capabilities of each tool in terms of detecting deepfakes. For example, the F1-Score shows how well a tool can find fake content while avoiding mistakes. These metrics give a clear picture of a tool's quality. Unlike selection criteria, evaluation metrics focus on how well the tool works. By using these metrics, we can clearly see which tools perform best and where they

[1]The class imbalance problem typically occurs when there are many more instances of some classes than others [57].

Table 3.2: Evaluation Metrics

Evaluation Metrics	**Description**
Processing time and Scalability	Evaluates the tool's speed in detecting deepfakes and its performance with larger or multiple inputs.
Interpretability	Evaluates the clarity and transparency of the tool's output, ensuring users can understand the detection results.
Detection Accuracy	Measures the correct identification of both genuine and deepfake content in the dataset, encompassing true positives and true negatives.
Precision	Measures the tool's capacity to avoid false positives by calculating the proportion of true positive results in the total predicted positives, including both true and false positives.
Recall	Measures the tool's accuracy in identifying actual deepfakes among all presented genuine content by calculating the number of true positives it identifies.
F1-Score	The harmonic mean of Precision and Recall. It provides a balance between the two when there's an uneven class distribution.
Dataset Compatibility	The tool's adaptability was checked by its dataset compatibility. Versatile tools or those recommended for specific datasets were preferred.
Training Models	The training models and techniques used by each tool were examined.
Face Detection Techniques	The face detection techniques and alternative methods used by the tools were assessed.
Programming Language and Framework	The programming languages and frameworks used in the tool's development were analyzed for insights into its capabilities.
Potential for Further Developempent	The potential for tool enhancement was assessed, along with any provided guidelines for such improvements.

might have problems. This makes it easier for researchers and users to choose the right tool for their needs. An overview of the evaluation metrics used in this study are provided in Table 3.2.

3.3 Datasets

Datasets are very important when working with deepfakes. They form the backbone of both the creation and detection of deepfakes. They also help us train models to create or spot deepfakes. Today, there are many datasets available that focus on both deepfake images and videos [82], [151], [78]. In Figure 3.1 a fake sample image of each dataset is provided.

3.3.1 FaceForensics++

FaceForensics++ (FF++) [2] serves as a broad dataset designed for forensic studies. It comprises 977 videos sourced from YouTube, over 1,000 unique sequences, and an impressive collection of more than 8,000 Deepfake videos [103], [109]. Originally launched in 2018, the dataset received significant updates in 2019 [109], making it richer and more diverse. As highlighted by the creators in their 2019 paper [103], the videos in the dataset were modified using a mix of techniques. This includes two graphics-driven methods, namely Face2Face [120] and FaceSwap [70], as well as two methods used in machine learning: DeepFakes [23] and Neural Textures [119]. A noteworthy feature of these manipulation methods is their reliance on both source and target video pairs for input. This makes the dataset a valuable resource for those looking to understand the nuances and intricacies of different deepfake generation techniques.

This dataset is offered in two quality versions: uncompressed and H264 [128] compressed format [82]. These versions allow for the assessment of deepfake detection methods on both compressed and uncompressed videos. However, the FF++ dataset struggles to accurately represent lip-sync deepfakes, and some videos display color discrepancies around the manipulated faces [82].

3.3.2 Deepfake Detection Challenge Dataset

Launched by Facebook community the Deepfake Detection Challenge (DFDC) dataset emerged from a collaborative initiative hosted on Kaggle [64], aiming to combat the rise of deceptive deepfake videos. Started in 2019 and ending with a big competition in 2020, this challenge saw over 2200 teams competing for an overall prize of one million

[2] https://github.com/ondyari/FaceForensics

dollars [109]. This challenge wasn't just about competing. It was a call for researchers all over the world to create new tools to spot fake content. The dataset has 104,500 different deepfake videos from 3,426 paid actors [29], [109]. This variety makes it a great tool to test and improve deepfake detection methods. The dataset is created using multiple face-swap techniques combined with different augmentations, such as geometric and color changes, and variations in frame rate. Additionally, it includes distractors like the incorporation of various objects within a video [82].

3.3.3 Face Forensics in the Wild

The Face Foresics in the Wild (FFIW)[3] dataset offers 10,000 high-quality manipulated videos [149]. What's unique about it is the automatic manipulation process. This process is managed by a domain-adversarial quality assessment network, which means creating this dataset requires less human intervention. This design ensures that the dataset can be scaled up easily and at a low human cost [149].

3.3.4 OpenForensics

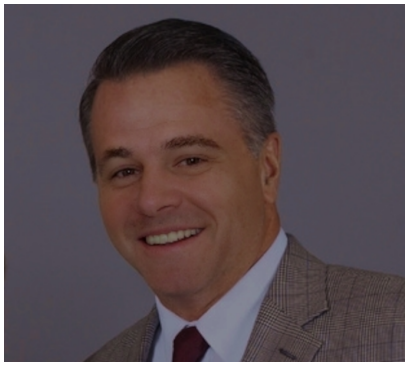
The OpenForensics[4] dataset is designed for detecting and segmenting multi-face forgeries. Its version 1.0.0 houses more than 115,000 real-world images, capturing a total of 334,000 human faces. Each image in the dataset comes with detailed face-related annotations, like the type of forgery, bounding boxes, segmentation masks, forgery boundaries, and typical facial landmarks [74].

[3] https://github.com/tfzhou/FFIW
[4] https://github.com/ltnghia/openforensics

(a) FaceForensics++

(b) DFDC

(c) FFIW



(d) OpenForensics

Figure 3.1: Sample deepfake images taken from FaceForensics++ [103], DFDC [29], FFIW [149] and OpenForensics [74].

4 Analysis of Publicly-Available Deepfake Detection Tools

For this study, several tools were selected for analysis. Their selection was not arbitrary; instead, it was based on the criteria detailed in Table 3.1. Tools were chosen with a focus on public availability, ensuring that everyone can access and benefit from them. Every tool in this study is open to the public, making the findings broadly applicable.

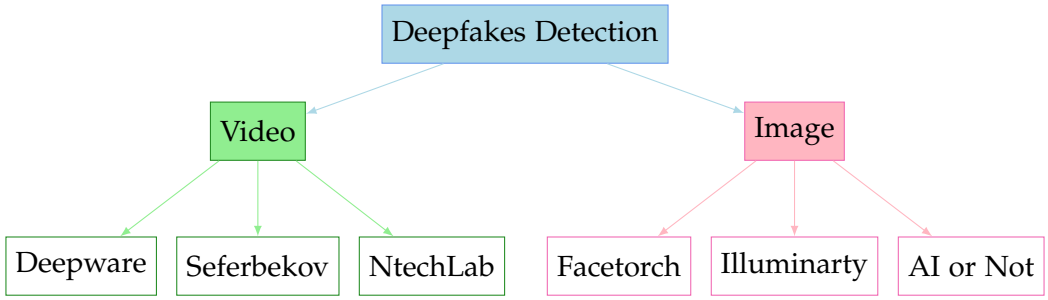


Figure 4.1: Categorization of deepfake detection tools

As depicted in Figure 4.1, three tools were chosen for video detection and another three for image detection. Every tool comes with its own strengths and weaknesses, ranging from how users can access it, the ease of installation, to understanding the results it produces. The selection aimed to cover a range of capabilities, as shown in Table 3.2, to ensure a general analysis.

For a general assessment, each one of these video detection tools were subjected to rigorous testing using a collection of 110 videos. Out of these, 80 videos were independently generated, 40 deepfakes using the DeepFaceLab (Subsection 2.2.1) and 40 deepfakes using FaceSwap (Subsection 2.2.2) tools. The remaining 30 were sourced from the established FaceForensics++ (Subsection 3.3.1) database, wherein 10 were identified as deepfakes, while 20 were genuine videos.

For the image detection analysis, a total of 123 images were used. Out of these, 103 were independently generated, 49 deepfakes produced using FaceApp (Subsection 2.2.5) and 54 deepfakes using Stable Diffusion (Subsection 2.2.3), while the remaining 20 were authentic images from our proprietary dataset [141].

4.1 Deepware

Deepware [25] is a tool made to tackle a big problem: the rise of fake videos. The people behind Deepware saw this challenge early on. Their parent company, Zemana[1], was looking into making AI tools for computer protection. However, by mid-2018, a pivot was observed, and the focus was redirected towards deepfake detection by the newly formed Deepware AI team [24].

A concern raised by Deepware's team is the potential advent of deceptive voice manipulations, which, when combined with video manipulations, could amplify the risks of scams and misinformation. This perspective highlights the urgency to develop reliable countermeasures against such threats.

Utilizing Deepware is simple. It offers an intuitive website [25] for uploading deepfakes for evaluation, and there's also an Android App [144] available for download and use. No advanced hardware requirements are demanded from users, making it accessible to many. The tool has a support team, and users have the freedom to test with datasets of their preference. Importantly, for the detection, Deepware employs 4 different deepfake detection models.

- **Avatarify**: Avatarify [6] is a tool that puts another person's face on yours during live video chats. It's available on Github for everyone [4]. Ali Aliev made Avatarify using a model from the University of Trento and Snap, Inc [17], [112]. This model can animate a photo with a video of someone else without needing many pictures of the face. Unlike other face-swap tools, this one works in real-time using similar facial features [43].

- **Deepware**: Deepware employs its proprietary detection models, but the source code isn't publicly accessible, limiting detailed information about its detection techniques.

- **Seferbekov**: The Seferbekov [110] model is employed as well. For in-depth details, refer to Section 4.2.

- **Ensemble**: Deepware utilizes ensemble learning, a method that merges multiple models to enhance accuracy and robustness. This ensemble approach combines simpler models, aiming for improved generalization [40]. Specifically, Deepware's ensemble combines its own model with the Seferbekov tool.

Besides detection capabilities, Deepware offers details about the analyzed video and audio inputs. For video, it displays Duration, Resolution, Frame Rate, and Codec,

[1] `https://zemana.com/us/antimalware.html`

while for audio (if present), it shows Duration, Channel, Sample Rate, and Codec. An overview is provided in Figure 4.2

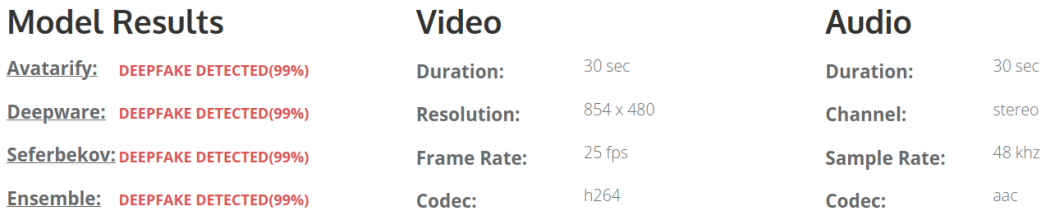


Figure 4.2: Overview of the capabilities of Deepware. Screenshot from [25]

A notable feature integrated into Deepware's platform allows for expert reviews. If mistakes in the deepfake detection process are suspected, a specialized review can be requested, underscoring the team's commitment to accuracy and continuous improvement. Deepware also provides a RESTful API[2] [8] and Software Development Kit (SDK) for developers looking to analyze videos from within their development setting. Privacy Policy[3] and Terms of Use[4] are also provided by Deepware.

The outcomes of the Detection Accuracy, Precision, Recall and F1-Score assessed with Deepware are presented in Table 4.1 and in Table 4.2.

Table 4.1: Computed data using Deepware for calculating evaluation metrics listed in Table 3.2

#Deepfakes	#Genuine videos	#True Positives	#True Negatives	#False Positives	#False Negatives
90	20	85	12	8	5

Table 4.2: Computed metrics using Deepware

Processing Time	Detection Accuracy	Precision	Recall	F1-Score
Avg. 21,5 sec	88,18%	91,40%	94,44%	92,9%

Detection Accuracy is the overall correct classification of the tool, considering both True Positives (TP) (deepfakes correctly identified) and True Negatives (TN) (genuine

[2] `https://api.deepware.ai`
[3] `https://deepware.ai/privacy-policy/`
[4] `https://deepware.ai/terms-of-services/`

videos correctly identified). It is calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{Total} \cdot 100\% \tag{4.1}$$

Precision is the proportion of videos correctly identified as deepfakes (TP) out of all instances (True Positives + False Positives (FP)) that the tool classified as deepfakes. It is calculated as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \cdot 100\% \tag{4.2}$$

Recall is also known as sensitivity, it is the proportion of True Positives in relation to the sum of True Positives and False Negatives (FN). It is calculated as follows:

$$\text{Recall} = \frac{TP}{TP + FN} \cdot 100\% \tag{4.3}$$

F1-Score is a metric providing balance between precision and recall, offering a more comprehensive view of the performance of the tool. It is calculated as follows:

$$\text{F1-Score} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} = \frac{TP}{TP + \frac{1}{2}(FN + FP)} \cdot 100\% \tag{4.4}$$

One potential drawback of Deepware is the apparent lack of updates since 2021. This could indicate that no new features or improvements have been introduced to the tool in recent years, potentially affecting its adaptability to newer deepfake techniques.

4.2 Seferbekov

Selim Seferbekov's [110] deepfake detection tool emerged as a winner in the DFDC challenge, securing a whopping prize of $500,000 [64]. Its acclaim is a testament to its advanced capabilities and effectiveness in identifying deepfakes.

At its core, Seferbekov's tool operates by examining videos frame-by-frame. In simpler terms, instead of looking at a video as one whole piece, it breaks it down and studies each frame just like individual pictures. This is essential because deepfakes can often differ in quality from one frame to another.

A major strength of this tool comes from its encoder, the EfficientNet B7 [117]. This encoder is like the brain of the tool and is recognized as one of the best of its kind. What makes it even more special is that it was trained using both ImageNet [104], a huge database of images, and a method called 'noisy student'. This 'noisy student' technique, as detailed in a research paper [140], allows the tool to learn better and improve its accuracy.

For preprocessing, face bounding boxes were extracted using the Multitask Cascaded Convolutional Neural Network (MTCNN) [147], a widely-used method for detecting faces and their landmarks. These bounding boxes helped determine crop positions, with a margin added around the face to better detect image noise differences. If multiple faces appeared in a video frame, the Structural Similarity Index Measure (SSIM) [125] was used to compare the fake and original frames. SSIM measures image differences based on luminance, contrast, and structure. Minimal differences could change a label from fake to real. Training was done with images sized 380×380 pixels in batches of 12 [109].

For each video it analyzes, Seferbekov's tool focuses on 32 specific frames. Rather than just averaging out the results from these frames, a unique method is employed to analyze them, which has proven to be quite effective. The tool uses five distinct EfficientNet B7 models, allowing it to analyze content from multiple perspectives.

However, to run this tool and train it, some powerful computer hardware is needed. It requires at least four GPUs that have a memory of 12gb or more. If someone is using popular graphics cards like the 1080Ti or 2080Ti, they might need to adjust some settings to get it working perfectly.

To utilize Seferbekov's detection tool, users need to download the detection models, add the deepfake videos to the tool's repository, and install certain Python libraries. In this study, Google Colaboratory was used to test the tool. The results of the metrics tested are presented in the table below. Furthermore, the tool is capable of analyzing several videos simultaneously. Simply group all your deepfake videos in a single folder and provide that folder as input.

Table 4.3: Computed data using Seferbekov's tool for calculating evaluation metrics listed in Table 3.2

#Deepfakes	#Genuine videos	#True Positives	#True Negatives	#False Positives	#False Negatives
90	20	79	18	2	11

Table 4.4: Computed metrics using Seferbekov's tool

Processing Time	Detection Accuracy	Precision	Recall	F1-Score
Avg. 67,25 sec	88,18%	97,53%	87,8%	92,41%

The calculations of the Detection Accuracy, Precision, Recall and F1-Score are provided in Section 4.1.

In terms of interpretability, Seferbekov's tool delivers simple outputs. Once the detection is complete, it provides users with a Comma-separated Values (CSV) file detailing the probability of a video being a deepfake. Its clear output format means users can quickly grasp the findings. Furthermore, the tool's ability to support numerous public deepfake datasets highlights its adaptability and relevance in the field of deepfake detection.

Additionally, Seferbekov's tool is open-source, providing an opportunity for those with a programming background and knowledge in detection techniques to modify and extend the implementation to their specific requirements. This flexibility encourages continuous improvement and adaptability to emerging deepfake trends.

One potential limitation of Seferbekov's tool is its lack of active support and community documentation. This means troubleshooting can be challenging if users encounter issues. Moreover, the tool's GitHub repository hasn't seen updates since 2021. Consequently, recent advancements in deepfake techniques might not be as effectively addressed by this detection tool, potentially impacting its utility with newer deepfake methods.

4.3 NtechLab

NtechLab's [94] software made waves by clinching third place in the DFDC Challenge, walking away with a cool $100,000 prize [64]. One of the standout features of this tool is its ability to check multiple deepfake videos simultaneously. After the analysis, it gives results in a straightforward CSV file, detailing the chances of videos being manipulated.

Before starting with this tool, the necessary training models had to be downloaded. It should be noted that a decent hardware configuration is essential for the tool to operate without issues.

While the tool's developers used superior hardware, the decision was made to work with Google Colaboratory in our research due to the lack of access to such advanced hardware. This choice highlights the adaptability of the tool across various platforms.

Talking about the technicalities on how this tool works: the heart of the software is based on a trio of EfficientNet-B7 models [117]. These models use Noisy Student [140] pre-trained weights. One of these models looks at video sequences, while the other two break videos down, frame by frame, changing their approach based on how big the face in the video is and a few other tweaks.

To make sure the models are spot-on and don't get confused or overdo things, a special mixup technique was used, alongside some neat tweaks like AutoAugment [21], Random Erasing [148] and various video compression parameters [94]. During training,

video compression was applied in real-time. Short cropped sequences, each containing 50 frames, were stored in PNG format. In every training cycle, these sequences were re-encoded with random settings using ffmpeg [130]. Because of the mixup, the model predictions were 'uncertain'. To enhance the model confidence during inference, a basic transformation was applied. The final decision was derived by averaging the model's predictions, with weights based on confidence. The combined time for training and preprocessing was roughly 5 days on DGX-1 [134].

The NtechLab's tool is open source, which means anyone can access and modify its code. This allows for greater flexibility, especially for those who wish to customize or build upon the tool's features. Its open nature encourages collaboration and adaptation to suit various needs.

The outcomes of the Detection Accuracy, Precision, Recall and F1-Score assessed with NtechLab's tool are presented in Table 4.5 and Table 4.6.

Table 4.5: Computed data using NtechLab's tool for calculating evaluation metrics listed in Table 3.2

#Deepfakes	#Genuine videos	#True Positives	#True Negatives	#False Positives	#False Negatives
90	20	79	19	1	11

Table 4.6: Computed metrics using NtechLab's tool

Processing Time	Detection Accuracy	Precision	Recall	F1-Score
Avg. 8 min	98%	98,75%	87,8%	93%

A shared concern with both NtechLab's and Seferbekov's tools is the lack of continuous support and detailed documentation. While Seferbekov's repository has remained inactive since 2021, NtechLab's hasn't seen updates since 2020. This stagnation hints at potential challenges in adapting to the latest deepfake techniques and staying current in the rapidly advancing realm of deepfake detection.

4.4 Facetorch

Facetorch [33], crafted by Tomáš Gajarský, stands out as a tool designed for image forgery detection. It is developed in Python and PyTorch[5], a popular programming framework. It is designed to identify faces and explore the detailed aspects of facial

[5] `https://pytorch.org/`

features using specific algorithms known as neural networks. The aim is to bring together the finest pre-existing models, enhance their speed with TorchScript[6], and present an all-in-one solution for face analysis. The features it offers include:

- Face Detection [27]
- Facial Representation Learning [11]
- Face Verification [63], [66]
- Facial Expression Recognition [108]
- Deepfake Detection [81], [110]
- 3D Face Alignment [139]

Notably, this tool comes with a user manual [37], API instructions [34], and an instance on Hugging Face [36], where users benefit from a dedicated interface for deepfake testing. Additionally, for those familiar with Google Colab, there's a ready-to-use notebook available [35]. One of its most appealing aspects is its open-source nature, with updates being consistently rolled out up until March 2023 [33].

The results of the Detection Accuracy, Precision, Recall and F1-Score assessed with Facetorch are presented in Table 4.7 and Table 4.8.

Table 4.7: Computed data using Facetorch for calculating evaluation metrics listed in Table 3.2

#Deepfakes	#Genuine videos	#True Positives	#True Negatives	#False Positives	#False Negatives
101*	20	2	20	0	99

* - Two images couldn't be detected but the overall number of deepfakes is 103 as mentioned in the last paragraph of Chapter 4.

Table 4.8: Computed metrics using Facetorch

Processing Time	Detection Accuracy	Precision	Recall	F1-Score
Avg. 13,5sec	18,18%	100%	1,98%	3,88%

Based on the results, it's clear that Facetorch had some challenges spotting deepfakes. One possible explanation is that the way these deepfakes were created might not match well with the methods the tool uses. Even with this shortcoming in detecting fake

[6] `https://pytorch.org/docs/stable/jit.html`

images, the tool still offers valuable information. Not only does it tell us if an image is real or fake, but it also helps identify the emotions shown in the image, like whether the person is feeling happy, angry and so on.

4.5 Illuminarty

Illuminarty [54] is an online platform where users can easily upload images to detect deepfakes and AI generated images [121]. It comes with an API option and offers subscription-based plans for those who want to enable more features such as: AI model identification for image generators or unlimited API usage. Using the free version, users can only check AI and Deepfake images and texts. There is also a user-friendly Terms of Use[7] and Privacy Policy[8] sections that explain the tool's purpose and usage instructions. Additionally, if users need suport, they can reach out the Illuminarty community through Discord [52] or Patreon [53].

The results of the Detection Accuracy, Precision, Recall and F1-Score assessed with Illuminarty are presented in Table 4.7 and Table 4.8.

Table 4.9: Computed data using Illuminarty for calculating evaluation metrics listed in Table 3.2

#Deepfakes	#Genuine videos	#True Positives	#True Negatives	#False Positives	#False Negatives
103	20	26	17	3	77

Table 4.10: Computed metrics using Illuminarty

Processing Time	Detection Accuracy	Precision	Recall	F1-Score
Avg. 3,9sec	35%	89,7%	25,24%	39,4%

4.6 AI or Not

AI or Not [97] is the idea of Andrey Doronichev, formerly Director of Product at Google. In 2022 he founded Optic[9], a startup dedicated to identifying the authenticity of images, videos and voices. Optic offers three main products:

[7] urlhttps://illuminarty.ai/en/terms.html
[8] `https://illuminarty.ai/en/privacy.html`
[9] `https://www.optic.xyz/`

AI or Not: This tool helps to determine whether an image has been generated by AI or a human.

Bias-o-Meter: This is a smart tool, integrated with the Chrome extension to find hidden biases and truth behind the News.

NFT fraud detection: This is tool prevents digital art or Non-Fungible Token (NFT) fraud across blockchains and marketplaces using realtime detection.

We will be focusing on AI or Not. This tool stands out because of its support and user-friendly guidelines. To analyze the authenticity of images, you can either upload an image directly to their website or input an image's web address. While it isn't purely for detecting deepfakes, it's versatile enough to identify images generated by technologies like Stable Diffusion, MidJourney[10], DALL-E[11] and GAN [28].

The results of the Detection Accuracy, Precision, Recall and F1-Score assessed with AI or Not are presented in Table 4.11 and Table 4.12.

Table 4.11: Computed data using AI or Not for calculating evaluation metrics listed in Table 3.2

#Deepfakes	#Genuine videos	#True Positives	#True Negatives	#False Positives	#False Negatives
100*	20	49	20	0	51

* - Three images couldn't be detected but the overall number of deepfakes is 103 as mentioned in the last paragraph of Chapter 4.

Table 4.12: Computed metrics using AI or Not

Processing Time	Detection Accuracy	Precision	Recall	F1-Score
Avg. 3,2sec	57,5%	100%	49%	65,77%

It also offers Chrome extension and a Telegram bot[12], if you want to stay updated. Your uploaded images are kept private; they've detailed this in their privacy policy. If you need to test multiple images at once, the documentation guides you on how to proceed after acquiring API access. The tool comfortably supports popular image formats like JPEG and PNG. Privacy Policy[13] and Terms of Use[14] are also provided by Deepware.

[10] https://www.midjourney.com/
[11] https://openai.com/dall-e-2
[12] https://t.me/AI_or_not_bot
[13] https://www.aiornot.com/privacy-policy
[14] https://www.aiornot.com/terms-of-service

Regarding the interpretability of the output, it tells you if the image is generated by AI or Human. In the words of the creators, AI or Not is a special online tool that quickly tells you if an image is made by a computer or a person. If the image is made by a computer, the tool even tells you the exact AI method used.

5 Case Studies

With the bond of art and technology, deepfakes are slowly redifing the state of entertainment. Their ability to transform audios and visuals offers creators better possibilities to create new type of content. From refining the quality of amateur videos to colorizing black and white movies, deepfakes are reshaping the entertainment and art industries.

5.1 Entertainment and Art

Deepfakes are popular in many creative areas. For example, a rapper Kendrick Lamar, used deepfake in 2022 music video to take on looks of famous celebrities. In the renewed Star Wars series, deepfake technology was used to resurrect characters like Princess Leia and Moff Tarkin, despite the original actors having passed away [87].

The real question is: Is deepfake technology a blessing or a curse for the talent? Of course it offers scalability. An actor can feature in global commercials or websites without constant traveling or learning new languages. For instance, Synthesia[1] did this with two commercials starring rapper Snoop Dogg. Insead of reshooting for a rebranded commercial, they altered Snoop Dogg's mouth movements to match the new brand name using deepfakes [73].

One of the positive implications of deepfakes in Arts industry is for example, the Salvador Dalí Museum introduced *Dalí Lives*, a digital revival of the deceased artist Salvador Dalí using deepfakes [127], [107]. This allows visitors to interact with the artist, hearing tales from his life, and even taking selfies. The Museum used an encoder-decoder deepfake technique, training encoders on Dalí's images and footage. An actor resembling Dalí was then mapped with Dalí's features using decoders (Figure 5.1).

Deepfakes, while being helpful and revolutionary, have notable weaknesses and can pose serious threats. They have been misused for creating fake celebrity videos, commiting fraud, and manipulating political content, leading California to ban making political deepfakes during election season in 2019 [127], [75].

[1] `https://www.synthesia.io/`

Figure 5.1: Screenshot taken from Dalí Lives [118].

5.2 Politics and Media

Deepfakes have had an influence on politics and media as well. On positive side of this context, some political figures have used deepfakes in their campaigns for creative advertisements. For instance, during the 2020 Delhi Legislative Assembly election in India, a deepfake video of the president of India's Bharatiya Janata Party (BJP) party, Manoj Tiwari, spread on WhatsApp, as reported by Vice [16]. In this first-ot-its-kind campaign use, the original video of Tiwari speaking English was changed to appear as if he spoke in Haryanvi, a Hindi dialect, targeting specific voters. The BJP collaborated with The Ideaz Factory to produce such deepfakes, aiming to reach India's diverse linguistic audience. This particular deepfake reached reportedly 15 million people across WhatsApp groups [58].

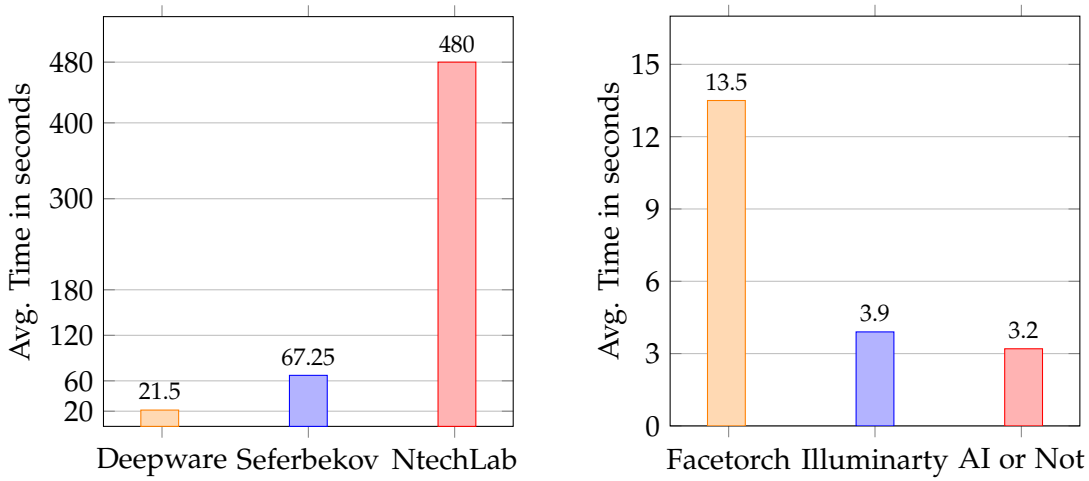
However, the implications of deepfakes in politcs and media are significat. The risk of spreading misinformation is high. There have been past occurrences where fake videos were used to damage the reputation of political individuals. Regulating political deepfakes is complex. While potential laws could aim to ban manipulated content of politicians, they'd need exceptions to safeguard artistic and satirical content, making implementation of those laws challenging due to the nature of satire and art [31], [68]. Promising solutions are emerging in tech industry, as tech giants like Adobe and Microsoft, alongside startups like Truepic, are developing tools for authenticity verification.

Ultimately, promoting digital media awareness is esstial, encouraging everyone to be critical of what they witness.

6 Results

In this chapter, the results of the in Chapter 4 computed experinments are presented. Responses to the research questions are addressed in Section 6.2. As previously described, three different video and image detection tools were used to detect deepfakes. Calculations for the video detection tools were based on the FaceForensics++ dataset, along with deepfake videos produced by DeepFaceLab and FaceSwap. Image detection tools were evaluated using images created by FaceApp and Stable Diffusion, as well as 20 authentic images.

6.1 Comparative Results of tools



(a) Video detection

(b) Image detection

Figure 6.1: Comparison of the Processing Time of detection tools from Chapter 4

From Figure 6.1, it's evident that image detection tools process data faster than the video detection tools. This discrepancy is due to the nature of the tools: image detection tools only process a single image (frame), whereas video detection tools may need to break a video into numerous frames and analyze each one separately, which

is considerably more time-consuming. Additionally, among the video detection tools, NtechLab's tool was slower than both Deepware and Seferbekov's tool. This might be because of the detection techniques they used.

The duration a tool takes to process data is tied to the size of the input. For instance, Seferbekov's tool, when handling a video under 10MB, averaged a processing time of 20 seconds. However, when confronted with a video exceeding 70MB, the time nearly tripled to almost a minute. The size-to-time relationship is consistent across all tools, meaning the larger the size of the input, the longer it's going to take the tool to process it.

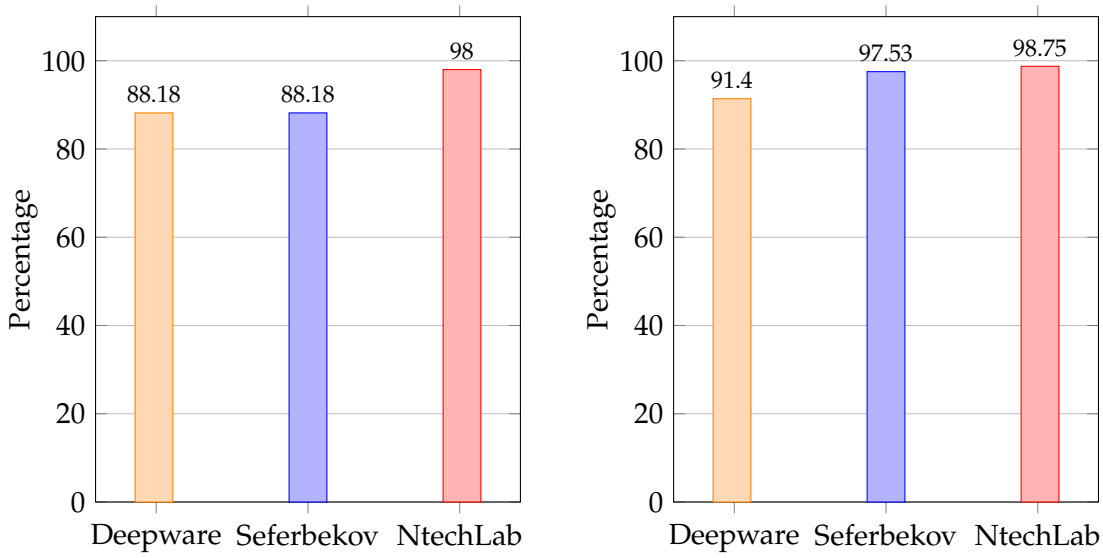
A comparison of assessed metrics for video detection tools is displayed in Figure 6.2. The comparison of which tool performed better is interesting. As we know, accuracy and precision are defined as follows: Accuracy measures the fraction of all instances that are correctly identified and precision measures the fraction of instances that were correctly predicted as positive out of all predicted positives. So precision is especially important when the number of false positives is high. In both of these cases NtechLab performed than the other two tools. This is due to the fact that NtechLab could detect more deepfakes.

Recall, also known as sensitivity, measures the proportion of actual positives that were correctly identified. It is crucial when the cost of missing a pisitive instance is high. And F-1 Score is a metric providing a balance between preicision and recall, offering a more comprehensive view of the performance. While NtechLab had higher accuracy and precision, Deepware and Seferbekov's tool have caught up in terms of Recall and F1-Score. This suggests that Deepware and Seferbekov's tool were more effective at correctly identifying true positive cases.

When it comes to image detection tools in Figure 6.3, AI or Not and Illuminarty outperformed Facetorch. One possible reason could be that AI or Not and Illuminarty are supported by companies and communities, receiving consisten updates. On the other hand, Facetorch is an open-source project. It hasn't had any updates since March 2023, which might make it less equipped to handle newer deepfake generation techniques.

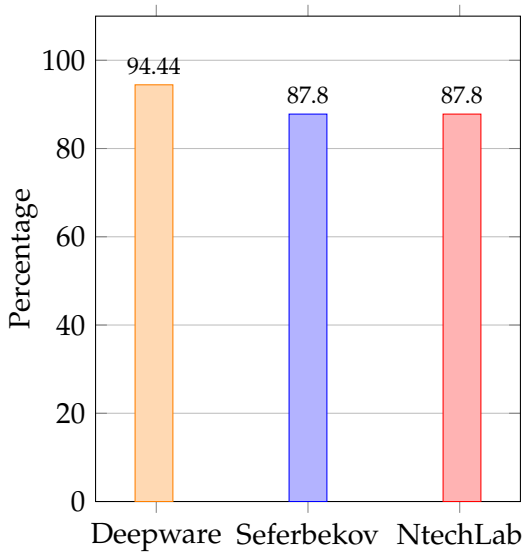
6.2 Final Results

To conclude the achieved results, it's noticable that some tools might have a high accuracy rate but perform poorly in terms of recall and F1-Score. Contrarily, certain tools might demonstrate high precision but low recall, which might indicate that the tool misses some actual positives. By comparing these metrics across the testes tools, a clearer insight into their strengths and limitations can be gained. This, in turn, helps us

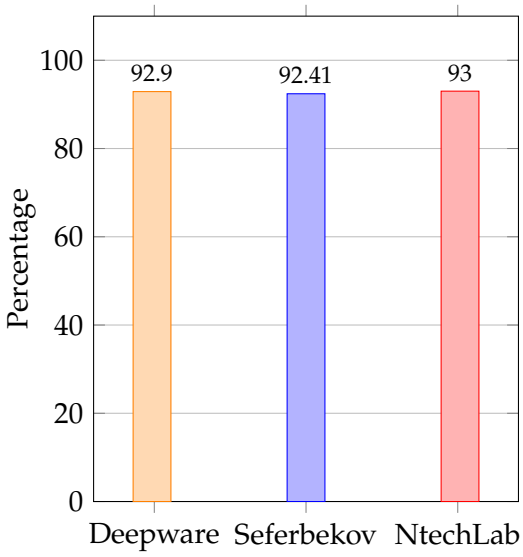


(a) Detection accuracy

(b) Precision

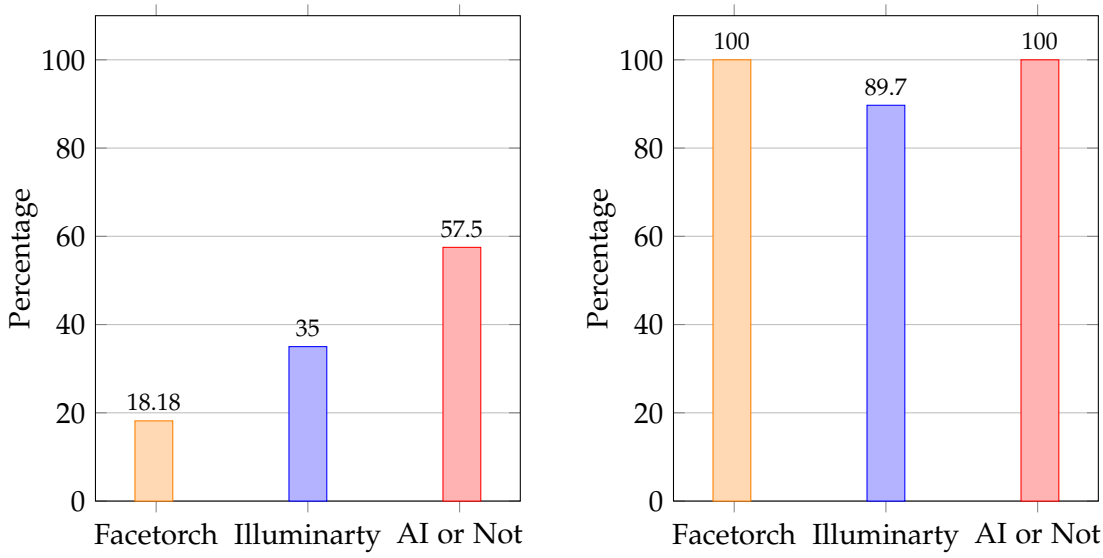


(c) Recall



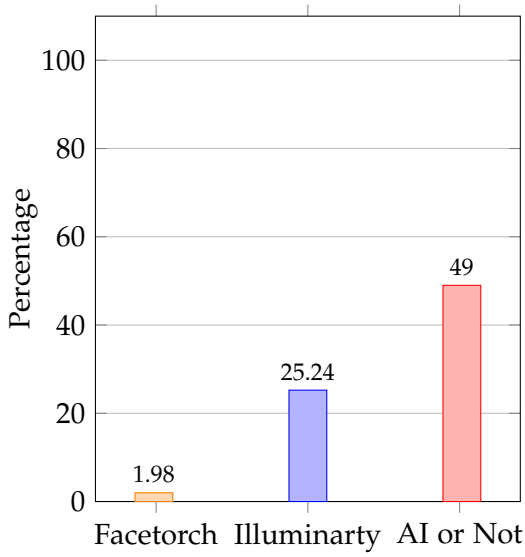
(d) F1-Score

Figure 6.2: Comparison of video detection tools from Chapter 4 according to the evaluation metrics listed in Table 3.2

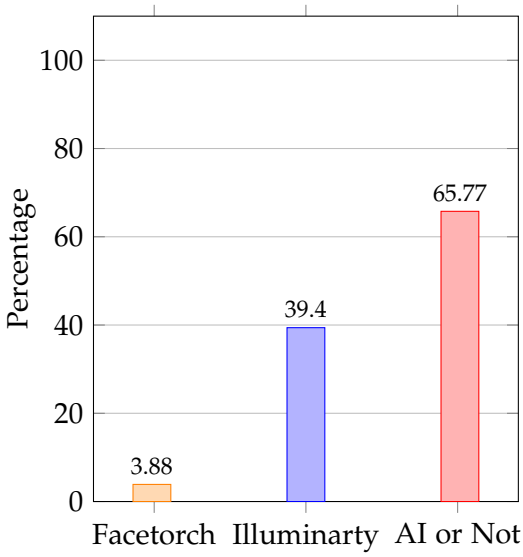


(a) Detection accuracy

(b) Precision



(c) Recall



(d) F1-Score

Figure 6.3: Comparison of image detection tools from Chapter 4 according to the evaluation metrics listed in Table 3.2

make important decisions on which tool is ideal for a particular task.

7 Discussion and Conclusion

7.1 Summary

In this research, it is demonstrated that the capabilities of deepfakes pose both opportunities for creative fields, technology advancements and threats to information integrity. Recognizing the increasing importance of this technology, this thesis analyzes various detection tools and their ability to counter manipulations. A primary focus is placed on video detection tools such as Deepware, Seferbekov's and NtechLab's tools. Through comprehensive testing, it was observed that while some tools showcased high Detection Accuracy and Precision, they might lag behind in other crucial metrics like Recall and F1-Score. These inconsistencies are important for a various evaluation approach, considering not just the accuracy but the tool's ability to capture true positives.

Moreover, image detection tools, including AI or Not, Illuminarty and Facetorch, were also observed. While AI or Not and Illuminarty emerged as the more proficient tool, possibly due to their consistent backing by dedicated companies and communities. Even though Facetorch showed potential as an open-source project, it fell behind, unserscoring the importance of regular updates to keep up with the changing landscape of deepfake techniques.

The diverse metrics used for assessment - Accuracy, Precision, Recall and F1-Score - illuminated that no single tool was generally superior in all fronts. Instead, their effectiveness is circumstantial, and the optimal tool selection should occur with the specific requirements of a task.

7.2 Future Work

There are several applicable areas to focus on for future work. One essential area of exploration is the expansion of datasets. By testing the tools against diverse and varied datasets, broader understanding of the tools can be gained. Moreover, testing with a larger volume of inputs can offer deeper insights into their scalability, robustness, and average performance metrics across datasets. Additionally, the need for regular updates and maintenance cannot be understated. As highlighted by the performance of tools such as Facetorch, staying updated is crucial to effectively tackle the latest deepfake

generation techniques. It would be also beneficial to not only keep the detection tools updated with the latest versions of these algorithms but also to retrain them periodically with updated techniques. By doing so, these tools can leverage the most recent advancements in the field. Lastly, a proactive studying the latest advancements in deepfake generation tools, allowing researchers develop detection methods based on new innovations, is crucial to this field.

7.3 Conclusion

Even though there's extensive research and numerous competitions focused on deepfake detection, no single method can identify them all. The swift advancements in deepfake generation could be a reason behind this. Since no approach has consistently shown to outpace the deepfake generator in effectiveness, it suggests that the world of deepfakes continues to develop. If it reaches a point where detection tools can't keep up, distinguishing between authentic and manipulated content might become nearly impossible. There's also a concern that if deepfake creators use detection tools as standards, it might unintentionally improve the quality of fake content. The work documented in this thesis is just a starting point among the countless opportunities that this field awaits.

Abbreviations

AI Artificial Intelligence

GAN Generative Adversarial Networks

ML Machine Learning

VAE Variational Autoencoders

CNN Convolutional Neural Networks

SD Stable Diffusion

IDE Integrated Development Environment

DFDC Deepfake Detection Challenge

FFIW Face Foresics in the Wild

TP True Positives

TN True Negatives

FP False Positives

FN False Negatives

GPU Graphics Processing Unit

CSV Comma-separated Values

API Application Programming Interface

BJP Bharatiya Janata Party

NFT Non-Fungible Token

RQ Research Questions

ML Machine Learning

MLP Multiplayer Perceptron Networks

FF++ FaceForensics++

SDK Software Development Kit

MTCNN Multitask Cascaded Convolutional Neural Network

SSIM Structural Similarity Index Measure

List of Figures

1.1 Deepfake of Bill Hader impersonating Arnold Schwarzenegger. Screenshot from [20] 1

2.1 Autoencoders common structure. Figure from [152] 9
2.2 Deepfake of Ivanka Trump impersonating Emma Watson. Screenshot from our own generated deepfake video dataset with DeepFaceLab. . . 12
2.3 Deepfake of Emma Stone impersonating Scarlett Johansson using Faceswap. Screenshot from [22]. 13
2.4 Deepfake of Justin Trudeau created with Stable Diffusion. Screenshot from our own generated deepfake video dataset with Stable Diffuison. 14
2.5 The reenactment synthesis process uses expression transfer to generate a UV map of the target actor that reflects the source actor's expression. This map, along with a background image, is processed by a neural renderer to create the final reenactment. Expression alteration is achieved by training a unique neural texture and renderer for the target actor, resulting in a manipulated video as shown in the image from [119]. . . 15
2.6 Overview of neural rendering pipeline [119]. 16
2.7 Deepfake created with FaceApp. Screenshot from our own generated deepfake image dataset with FaceApp. 17

3.1 Sample deepfake images taken from FaceForensics++ [103], DFDC [29], FFIW [149] and OpenForensics [74]. 27

4.1 Categorization of deepfake detection tools 28
4.2 Overview of the capabilities of Deepware. Screenshot from [25] 30

5.1 Screenshot taken from Dalí Lives [118]. 40

6.1 Comparison of the Processing Time of detection tools from Chapter 4 . 41
6.2 Comparison of video detection tools from Chapter 4 according to the evaluation metrics listed in Table 3.2 43
6.3 Comparison of image detection tools from Chapter 4 according to the evaluation metrics listed in Table 3.2 44

List of Tables

2.1 Some of GANs with alternate architectures [9], [131] 11

3.1 Selection Criteria 22
3.2 Evaluation Metrics 24

4.1 Computed data using Deepware for calculating evaluation metrics listed in Table 3.2 30
4.2 Computed metrics using Deepware 30
4.3 Computed data using Seferbekov's tool for calculating evaluation metrics listed in Table 3.2 32
4.4 Computed metrics using Seferbekov's tool 32
4.5 Computed data using NtechLab's tool for calculating evaluation metrics listed in Table 3.2 34
4.6 Computed metrics using NtechLab's tool 34
4.7 Computed data using Facetorch for calculating evaluation metrics listed in Table 3.2 35
4.8 Computed metrics using Facetorch 35
4.9 Computed data using Illuminarty for calculating evaluation metrics listed in Table 3.2 36
4.10 Computed metrics using Illuminarty 36
4.11 Computed data using AI or Not for calculating evaluation metrics listed in Table 3.2 37
4.12 Computed metrics using AI or Not 37

Bibliography

[1] S. Agarwal and H. Farid. "Detecting Deep-Fake Videos From Aural and Oral Dynamics." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2021, pp. 981–989.

[2] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li. "Protecting World Leaders Against Deep Fakes." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2019.

[3] M. Albahar and J. Almalki. "Deepfakes: Threats and countermeasures systematic review." In: *Journal of Theoretical and Applied Information Technology* 97.22 (2019), pp. 3242–3250.

[4] A. Aliev. *Avatarify GitHub Repository*. Accessed: 29.07.2023. URL: https://github.com/alievk/avatarify-python.

[5] A. Amazon. *What is data labeling?* Accessed: 13.08.2023. URL: https://aws.amazon.com/sagemaker/data-labeling/what-is-data-labeling/.

[6] Avatarify. *Avatarify Website*. Accessed: 14.08.2023. URL: https://avatarify.ai/.

[7] P. Baheti. *Supervised and Unsupervised Learning [Differences & Examples]*. Accessed: 13.08.2023. Oct. 2021. URL: https://www.v7labs.com/blog/supervised-vs-unsupervised-learning.

[8] B. J. Bernard. "Deepfake Detection Framework." In: (). URL: https://excel.fit.vutbr.cz/submissions/2023/033/33.pdf.

[9] G. Boesch. *Guide to Generative Adversarial Networks (GANs) in 2023*. Accessed: 13.08.2023. 2023. URL: https://viso.ai/deep-learning/generative-adversarial-networks-gan/.

[10] A. Brock, J. Donahue, and K. Simonyan. *Large Scale GAN Training for High Fidelity Natural Image Synthesis*. 2019. arXiv: 1809.11096 [cs.LG].

[11] A. Bulat, S. Cheng, J. Yang, A. Garbett, E. Sanchez, and G. Tzimiropoulos. *Pre-training strategies and datasets for facial representation learning*. 2022. arXiv: 2103.16554 [cs.CV].

[12] A. Bulat and G. Tzimiropoulos. "How Far are We from Solving the 2D & 3D Face Alignment Problem? (and a Dataset of 230,000 3D Facial Landmarks)." In: *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2017. DOI: 10.1109/iccv.2017.116. URL: https://doi.org/10.1109%2Ficcv.2017.116.

[13] BuzzFeed. *You Won't Believe What Obama Says In This Video!* Accessed: 21.05.2023. Apr. 2018. URL: https://www.youtube.com/watch?v=cQ54GDm1eL0.

[14] D. Chakraborty. *IN DEPTH OF Faceapp*. Accessed: 14.08.2023. Apr. 2020. URL: https://medium.com/analytics-vidhya/in-depth-of-faceapp-a08be9fe86f6.

[15] B. Chesney and D. Citron. "Deep fakes: A looming challenge for privacy, democracy, and national security." In: *Calif. L. Rev.* 107 (2019), p. 1753.

[16] N. Christopher. *We've Just Seen the First Use of Deepfakes in an Indian Election Campaign*. Accessed: 10.08.2023. Feb. 2020. URL: https://www.vice.com/en/article/jgedjb/the-first-use-of-deepfakes-in-indian-election-by-bjp.

[17] S. Cole. *This Open-Source Program Deepfakes You During Zoom Meetings, in Real Time*. Accessed: 14.08.2023. Apr. 2020. URL: https://www.vice.com/en/article/g5xagy/this-open-source-program-deepfakes-you-during-zoom-meetings-in-real-time.

[18] CompVis. *Stable Diffusion Repository on GitHub*. Accessed: 13.08.2023. URL: https://github.com/CompVis/stable-diffusion.

[19] D. Cozzolino, G. Poggi, and L. Verdoliva. *Recasting Residual-based Local Descriptors as Convolutional Neural Networks: an Application to Image Forgery Detection*. 2017. arXiv: 1703.04615 [cs.CV].

[20] Ctrl Shift Face. *Bill Hader impersonates Arnold Schwarzenegger [DeepFake]*. Accessed: 13.07.2023. May 2019. URL: https://www.youtube.com/watch?v=bPhUhypV27w.

[21] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le. *AutoAugment: Learning Augmentation Policies from Data*. 2019. arXiv: 1805.09501 [cs.CV].

[22] Dailymotion. *Faceswap Phaze-A - 256px Demo*. Accessed: 05.08.2023. May 2021. URL: https://dai.ly/x810mot.

[23] Deepfakes. *FaceSwap: Deepfakes spftware for all*. URL: https://github.com/deepfakes/faceswap.

[24] Deepware. *Deepware About Us Section*. Accessed: 20.06.2023. URL: https://deepware.ai/about/.

[25] Deepware. *Deepware Website*. Accessed: 20.06.2023. URL: https://deepware.ai/.

[26] J. Delua. *Supervised vs. Unsupervised Learning: What's the Difference?* Accessed: 13.08.2023. Mar. 2021. URL: https://www.ibm.com/blog/supervised-vs-unsupervised-learning/.

[27] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou. "RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020.

[28] D. Djudjic. *BYE-BYE FAKE NEWS: THIS FREE TOOL SPOTS AI-GENERATED IMAGES IN A SECOND*. Accessed: 01.08.2023. June 2023. URL: https://www.diyphotography.net/optic-tool-spots-ai-generated-images/.

[29] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer. *The DeepFake Detection Challenge (DFDC) Dataset*. 2020. arXiv: 2006.07397 [cs.CV].

[30] R. Durall, M. Keuper, F.-J. Pfreundt, and J. Keuper. *Unmasking DeepFakes with simple Features*. 2020. arXiv: 1911.00686 [cs.LG].

[31] K. Farish. *Political Deepfakes: social media trend or genuine threat?* Accessed: 10.08.2023. Sept. 2022. URL: https://www.dacbeachcroft.com/en/gb/articles/2022/september/political-deepfakes-social-media-trend-or-genuine-threat/.

[32] M. Fisher. *How I Became the Fake Tom Cruise*. Accessed: 21.05.2023. July 2022. URL: https://www.hollywoodreporter.com/feature/deepfake-tom-cruise-miles-fisher-1235182932/.

[33] T. Gajarský. *Facetorch - GitHub Repository*. Accessed: 25.07.2023. URL: https://github.com/tomas-gajarsky/facetorch.

[34] T. Gajarský. *Facetorch Documentation*. Accessed: 29.07.2023. 2022. URL: https://tomas-gajarsky.github.io/facetorch/facetorch/index.html.

[35] T. Gajarský. *Facetorch Google Colab notebook*. Accessed: 29.07.2023. 2022. URL: https://colab.research.google.com/github/tomas-gajarsky/facetorch/blob/main/notebooks/facetorch_notebook_demo.ipynb.

[36] T. Gajarský. *Facetorch Hugging Face instance*. Accessed: 29.07.2023. 2022. URL: https://huggingface.co/spaces/tomas-gajarsky/facetorch-app.

[37] T. Gajarský. *Facetorch User Guide*. Accessed: 29.07.2023. Sept. 2022. URL: https://medium.com/@gajarsky.tomas/facetorch-user-guide-a0e9fd2a5552.

[38] D. Gamage, P. Ghasiya, V. Bonagiri, M. E. Whiting, and K. Sasahara. "Are Deepfakes Concerning? Analyzing Conversations of Deepfakes on Reddit and Exploring Societal Implications." In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3517446. URL: https://doi.org/10.1145/3491102.3517446.

[39] N. Gardiner. "Facial re-enactment, speech synthesis and the rise of the Deepfake." In: 2019. URL: https://api.semanticscholar.org/CorpusID:132624704.

[40] N. Giatsoglou, S. Papadopoulos, and I. Kompatsiaris. *Investigation of ensemble methods for the detection of deepfake face manipulations*. 2023. arXiv: 2304.07395 [cs.CV].

[41] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML].

[42] C. Goyal. *2023's Best Guide to Discriminative & Generative Machine Learning Models*. Accessed: 13.08.2023. June 2023. URL: https://www.analyticsvidhya.com/blog/2021/07/deep-understanding-of-discriminative-and-generative-models-in-machine-learning/#:~:text=In%20simple%20words%2C%20a%20discriminative,probability%20for%20a%20given%20example..

[43] T. Greene. *Watch: Fake Elon Musk Zoom-bombs meeting using real-time Deepfake AI*. Accessed: 14.08.2023. Apr. 2020. URL: https://thenextweb.com/news/watch-fake-elon-musk-zoom-bombs-meeting-using-real-time-deepfake-ai.

[44] S. Greengard. "Will Deepfakes Do Deep Damage?" In: *Commun. ACM* 63.1 (Dec. 2019), pp. 17–19. ISSN: 0001-0782. DOI: 10.1145/3371409. URL: https://doi.org/10.1145/3371409.

[45] L. Guilloux. *Swap faces on videos by means of AI*. Accessed: 13.08.2023. Mar. 2019. URL: https://www.malavida.com/en/soft/fakeapp/.

[46] Y. Guo, W. He, J. Zhu, and C. Li. "A Light Autoencoder Networks for Face Swapping." In: *Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence*. CSAI '18. Shenzhen, China: Association for Computing Machinery, 2018, pp. 459–462. ISBN: 9781450366069. DOI: 10.1145/3297156.3297210. URL: https://doi.org/10.1145/3297156.3297210.

[47] G. E. Hinton and R. R. Salakhutdinov. "Reducing the Dimensionality of Data with Neural Networks." In: *Science* 313.5786 (2006), pp. 504–507. DOI: 10.1126/science.1127647. eprint: https://www.science.org/doi/pdf/10.1126/

science.1127647. URL: https://www.science.org/doi/abs/10.1126/science.1127647.

[48] J. Ho and S. Ermon. *Generative Adversarial Imitation Learning*. 2016. arXiv: 1606.03476 [cs.LG].

[49] J. Ho, A. Jain, and P. Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. arXiv: 2006.11239 [cs.LG].

[50] huggingface.co. *Diffuse The Rest - a Hugging Face Space by huggingface*. Accessed: 13.08.2023. URL: https://huggingface.co/spaces/huggingface-projects/diffuse-the-rest.

[51] V. Iglovikov and A. Shvets. *TernausNet: U-Net with VGG11 Encoder Pre-Trained on ImageNet for Image Segmentation*. 2018. arXiv: 1801.05746 [cs.CV].

[52] Illuminarty. *Illuminarty Discord*. Accessed: 14.07.2023. URL: https://discord.gg/2GRUBCHduD.

[53] Illuminarty. *Illuminarty Patreon*. Accessed: 14.07.2023. URL: https://www.patreon.com/illuminarty.

[54] Illuminarty. *Illuminarty Website*. Accessed: 14.07.2023. URL: https://app.illuminarty.ai/.

[55] M. Inc. *Midjourney*. Accessed: 13.08.2023. URL: https://www.midjourney.com/home/?callbackUrl=%2Fapp%2F.

[56] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. "Image-to-Image Translation with Conditional Adversarial Networks." In: *CVPR* (2017).

[57] S. H. Javaheri, M. M. Sepehri, and B. Teimourpour. "Chapter 6 - Response Modeling in Direct Marketing: A Data Mining-Based Approach for Target Selection." In: *Data Mining Applications with R*. Ed. by Y. Zhao and Y. Cen. Boston: Academic Press, 2014, pp. 153–180. ISBN: 978-0-12-411511-8. DOI: https://doi.org/10.1016/B978-0-12-411511-8.00006-2. URL: https://www.sciencedirect.com/science/article/pii/B9780124115118000062.

[58] C. Jee. *An Indian politician is using deepfake technology to win new voters*. Accessed: 10.08.2023. Feb. 2020. URL: https://www.technologyreview.com/2020/02/19/868173/an-indian-politician-is-using-deepfakes-to-try-and-win-voters/.

[59] F. Jia and S. Yang. "Video face swap with DeepFaceLab." In: *International Conference on Computer Graphics, Artificial Intelligence, and Data Processing (ICCAID 2021)*. Ed. by F. Wu, J. Liu, and Y. Chen. Vol. 12168. International Society for Optics and Photonics. SPIE, 2022, 121681H. DOI: 10.1117/12.2631297. URL: https://doi.org/10.1117/12.2631297.

[60] S. Jia, X. Li, and S. Lyu. "Model Attribution of Face-Swap Deepfake Videos." In: *2022 IEEE International Conference on Image Processing (ICIP)*. 2022, pp. 2356–2360. DOI: 10.1109/ICIP46576.2022.9897972.

[61] E. Johansson. *Detecting Deepfakes and Forged Videos Using Deep Learning*. Accessed: 13.08.2023. 2020. URL: https://www.lunduniversity.lu.se/lup/publication/9019746.

[62] D. Johnson and A. Johnson. *What are deepfakes? How fake AI-powered audio and video warps our perception of reality*. Accessed: 13.07.2023. June 2023. URL: https://www.businessinsider.com/guides/tech/what-is-deepfake.

[63] J. Jung, S. Lee, H.-S. Oh, Y. Park, J. Park, and S. Son. *Unified Negative Pair Generation toward Well-discriminative Feature Space for Face Recognition*. 2022. arXiv: 2203.11593 [cs.CV].

[64] Kaggle. *Deepfake Detection Challenge*. Accessed: 27.07.2023. Dec. 2019. URL: https://www.kaggle.com/competitions/deepfake-detection-challenge/.

[65] E. Kan. *What The Heck Are VAE-GANs?* Accessed: 13.08.2023. Aug. 2018. URL: https://towardsdatascience.com/what-the-heck-are-vae-gans-17b86023588a.

[66] M. Kim, A. K. Jain, and X. Liu. *AdaFace: Quality Adaptive Margin for Face Recognition*. 2023. arXiv: 2204.00964 [cs.CV].

[67] D. P. Kingma and M. Welling. *Auto-Encoding Variational Bayes*. 2022. arXiv: 1312.6114 [stat.ML].

[68] C. Klein. *"This Will Be Dangerous in Elections": Political Media's Next Big Challenge Is Navigating AI Deepfakes*. Accessed: 10.08.2023. Mar. 2023. URL: https://www.vanityfair.com/news/2023/03/ai-2024-deepfake.

[69] I. Korshunova, W. Shi, J. Dambre, and L. Theis. "Fast Face-Swap Using Convolutional Neural Networks." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017.

[70] M. Kowalski. *FaceSwap GitHub Page*. URL: https://github.com/MarekKowalski/FaceSwap/.

[71] A. Kumar. *Generative vs Discriminative Models: Examples*. Accessed: 13.08.2023. Mar. 2023. URL: https://vitalflux.com/generative-vs-discriminative-models-examples/.

[72] K. Kurzhals. "Privacy in Eye Tracking Research with Stable Diffusion." In: *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications*. ETRA '23. Tubingen, Germany: Association for Computing Machinery, 2023. ISBN: 9798400701504. DOI: 10.1145/3588015.3589842. URL: https://doi.org/10.1145/3588015.3589842.

[73] V. Lalla, A. Mitrani, and Z. Harned. *Artificial intelligence: deepfakes in the entertainment industry*. Accessed: 02.08.2023. June 2022. URL: https://www.motionanalysis.com/biomechanics/deepfake-technology-for-entertainment-the-pros-and-cons/.

[74] T.-N. Le, H. H. Nguyen, J. Yamagishi, and I. Echizen. "OpenForensics: Large-Scale Challenging Dataset For Multi-Face Forgery Detection And Segmentation In-The-Wild." In: *International Conference on Computer Vision*. 2021.

[75] C. Lecher. *California has banned political deepfakes during election season*. Accessed: 09.08.2023. Oct. 2019. URL: https://www.theverge.com/2019/10/7/20902884/california-deepfake-political-ban-election-2020.

[76] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. *Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network*. 2017. arXiv: 1609.04802 [cs.CV].

[77] C. Leibowicz, J. Stray, and E. Saltz. *Manipulated Media Detection Requires More Than Tools: Community Insights on What's Needed*. Accessed: 13.07.2023. July 2020. URL: https://partnershiponai.org/manipulated-media-detection-requires-more-than-tools-community-insights-on-whats-needed/.

[78] G. Li, X. Zhao, Y. Cao, P. Pei, J. Li, and Z. Zhang. "FMFCC-V: An Asian Large-Scale Challenging Dataset for DeepFake Detection." In: *Proceedings of the 2022 ACM Workshop on Information Hiding and Multimedia Security*. IH&MMSec '22. Santa Barbara, CA, USA: Association for Computing Machinery, 2022, pp. 7–18. ISBN: 9781450393553. DOI: 10.1145/3531536.3532946. URL: https://doi.org/10.1145/3531536.3532946.

[79] Y. Li, M.-C. Chang, and S. Lyu. *In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking*. 2018. arXiv: 1806.02877 [cs.CV].

[80] F. T. Limited. *AceApp*. URL: https://www.faceapp.com/.

[81] C. Luo, S. Song, W. Xie, L. Shen, and H. Gunes. "Learning Multi-dimensional Edge Feature-based AU Relation Graph for Facial Action Unit Recognition." In: *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, July 2022. DOI: 10.24963/ijcai.2022/173. URL: https://doi.org/10.24963%2Fijcai.2022%2F173.

[82] M. Masood, M. Nawaz, K. Malik, A. Javed, A. Irtaza, and H. Malik. "Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward." In: *Applied Intelligence* 53 (June 2022), pp. 1–53. DOI: 10.1007/s10489-022-03766-z.

[83] P. Mehta, G. Jagatap, K. Gallagher, B. Timmerman, P. Deb, S. Garg, R. Greenstadt, and B. Dolan-Gavitt. "Can Deepfakes Be Created on a Whim?" In: *Companion Proceedings of the ACM Web Conference 2023*. WWW '23 Companion. Austin, TX, USA: Association for Computing Machinery, 2023, pp. 1324–1334. ISBN: 9781450394192. DOI: 10.1145/3543873.3587581. URL: https://doi.org/10.1145/3543873.3587581.

[84] R. Metz. *How a deepfake Tom Cruise on TikTok turned into a very real AI company*. Accessed: 12.08.2023. Aug. 2021. URL: https://edition.cnn.com/2021/08/06/tech/tom-cruise-deepfake-tiktok-company/index.html.

[85] Y. Mirsky and W. Lee. "The Creation and Detection of Deepfakes: A Survey." In: *ACM Comput. Surv.* 54.1 (Jan. 2021). ISSN: 0360-0300. DOI: 10.1145/3425780. URL: https://doi.org/10.1145/3425780.

[86] M. Mirza and S. Osindero. *Conditional Generative Adversarial Nets*. 2014. arXiv: 1411.1784 [cs.LG].

[87] Motion Analysis. *Deepfake technology for entertainment: the pros and cons*. Accessed: 02.08.2023. Aug. 2022. URL: https://www.motionanalysis.com/biomechanics/deepfake-technology-for-entertainment-the-pros-and-cons/.

[88] L. M. U. of Munich. *Revolutionizing image generation by AI: Turning text into images*. Accessed: 13.08.2023. Sept. 2022. URL: https://www.lmu.de/en/newsroom/news-overview/news/revolutionizing-image-generation-by-ai-turning-text-into-images.html.

[89] M. Mustak, J. Salminen, M. Mäntymäki, A. Rahman, and Y. K. Dwivedi. "Deepfakes: Deceptions, mitigations, and opportunities." In: *Journal of Business Research* 154 (2023), p. 113368. ISSN: 0148-2963. DOI: https://doi.org/10.1016/j.jbusres.2022.113368. URL: https://www.sciencedirect.com/science/article/pii/S0148296322008335.

[90] G. Nanos. *VAE Vs. GAN For Image Generation*. Accessed: 13.08.2023. Mar. 2023. URL: https://www.baeldung.com/cs/vae-vs-gan-image-generation.

[91] A. Neyaz, A. Kumar, S. Krishnan, J. Placker, and Q. Liu. "Security, Privacy and Steganographic Analysis of FaceApp and TikTok." In: (June 2020).

[92] H. H. Nguyen, J. Yamagishi, and I. Echizen. *Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos*. 2018. arXiv: 1810.11215 [cs.CV].

[93] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, and C. M. Nguyen. "Deep learning for deepfakes creation and detection: A survey." In: *Computer Vision and Image Understanding* 223 (Oct. 2022), p. 103525. DOI: 10.1016/j.cviu.2022.103525. URL: https://doi.org/10.1016%2Fj.cviu.2022.103525.

[94] NtechLab. *Deepfake Detection Challenge - GitHub Repository*. Accessed: 25.07.2023. URL: https://github.com/NTech-Lab/deepfake-detection-challenge.

[95] F. Offert and P. Bell. "Perceptual bias and technical metapictures: critical machine vision as a humanities challenge." In: *AI & SOCIETY* 36 (Dec. 2021), pp. 1–12. DOI: 10.1007/s00146-020-01058-z.

[96] OpenAI. *DALL·E 2 is an AI system that can create realistic images and art from a description in natural language.* Accessed: 13.08.2023. URL: https://openai.com/dall-e-2.

[97] Optic. *AI or Not - Website*. Accessed: 14.07.2023. URL: https://www.aiornot.com/.

[98] S. Parkin. *The rise of the deepfake and threat to democracy*. Accessed: 12.08.2023. June 2019. URL: https://www.theguardian.com/technology/ng-interactive/2019/jun/22/the-rise-of-the-deepfake-and-the-threat-to-democracy.

[99] I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, M. Dpfks, C. S. Facenheim, L. RP, J. Jiang, S. Zhang, P. Wu, B. Zhou, and W. Zhang. *DeepFaceLab: Integrated, flexible and extensible face-swapping framework*. 2021. arXiv: 2005.05535 [cs.CV].

[100] D. Pickell. *What Is FaceApp? The Technology Behind This AI-Enabled Mobile App*. Accessed: 14.08.2023. July 2019. URL: https://learn.g2.com/faceapp.

[101] A. Radford, L. Metz, and S. Chintala. *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*. 2016. arXiv: 1511.06434 [cs.LG].

[102] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2022. arXiv: 2112.10752 [cs.CV].

[103] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. "FaceForensics++: Learning to Detect Manipulated Facial Images." In: *International Conference on Computer Vision (ICCV)*. 2019.

[104] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. "ImageNet Large Scale Visual Recognition Challenge." In: *International Journal of Computer Vision (IJCV)* 115.3 (2015), pp. 211–252. DOI: 10.1007/s11263-015-0816-y.

[105] A. Sachdev. *Spatial and Frequency Domain — Image Processing*. Accessed: 10.08.2023. Oct. 2019. URL: https://medium.com/vithelper/spatial-and-frequency-domain-image-processing-83ffa3fc7cbc.

[106] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. *Improved Techniques for Training GANs*. 2016. arXiv: 1606.03498 [cs.LG].

[107] Salvador Dalí Museum. *dalí lives: museum brings artist back to life with ai*. Accessed: 09.08.2023. Jan. 2019. URL: https://thedali.org/press-room/dali-lives-museum-brings-artists-back-to-life-with-ai/.

[108] A. V. Savchenko. "Facial expression and attributes recognition based on multi-task learning of lightweight neural networks." In: *2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY)*. 2021, pp. 119–124. DOI: 10.1109/SISY52375.2021.9582508.

[109] M. Schicklgruber. *Master Thesis - Deepfake Detection*. Accessed: 10.06.2023. 2022. URL: https://epub.jku.at/obvulihs/download/pdf/8013601?originalFilename=true.

[110] S. Seferbekov. *Seferbekov GitHub Repository*. Accessed: 04.06.2023. URL: https://github.com/selimsef/dfdc_deepfake_challenge.

[111] H. F. Shahzad, F. Rustam, E. S. Flores, J. Luís Vidal Mazón, I. de la Torre Diez, and I. Ashraf. "A Review of Image Processing Techniques for Deepfakes." In: *Sensors* 22.12 (2022). ISSN: 1424-8220. DOI: 10.3390/s22124556. URL: https://www.mdpi.com/1424-8220/22/12/4556.

[112] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe. *First Order Motion Model for Image Animation*. 2020. arXiv: 2003.00196 [cs.CV].

[113] K. Simonyan and A. Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. arXiv: 1409.1556 [cs.CV].

[114] T. Smith. *Leaked deck raises questions over Stability AI's Series A pitch to investors*. Accessed: 13.08.2023. Apr. 2023. URL: https://sifted.eu/articles/stability-ai-fundraise-leak.

[115] stability.ai. *Stable Diffusion*. URL: https://stability.ai/stablediffusion.

[116] stability.ai. *Stable Diffusion Launch Announcement*. Accessed: 13.08.2023. Aug. 2022. URL: https://stability.ai/blog/stable-diffusion-announcement.

[117] M. Tan and Q. V. Le. *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. 2020. arXiv: 1905.11946 [cs.LG].

[118] The Dalí Museum. *Behind the Scenes, Dalí Lives*. Accessed: 09.08.2023. May 2019. URL: https://www.youtube.com/watch?v=BIDaxl4xqJ4.

[119] J. Thies, M. Zollhöfer, and M. Nießner. *Deferred Neural Rendering: Image Synthesis using Neural Textures*. 2019. arXiv: 1904.12356 [cs.CV].

[120] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. *Face2Face: Real-time Face Capture and Reenactment of RGB Videos*. 2020. arXiv: 2007.14808 [cs.CV].

[121] S. A. Thompson and T. Hsu. *How Easy Is It to Fool A.I.-Detection Tools?* Accessed: 14.07.2023. June 2023. URL: https://www.nytimes.com/interactive/2023/06/28/technology/ai-detection-midjourney-stable-diffusion-dalle.html.

[122] R. Toews. *Deepfakes Are Going To Wreak Havoc On Society. We Are Not Prepared*. Accessed: 12.08.2023. May 2020. URL: https://www.forbes.com/sites/robtoews/2020/05/25/deepfakes-are-going-to-wreak-havoc-on-society-we-are-not-prepared/?sh=4b68fb7a7494.

[123] C. Vaccari and A. Chadwick. "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News." In: *Social Media + Society* 6.1 (2020), p. 2056305120903408. DOI: 10.1177/2056305120903408. eprint: https://doi.org/10.1177/2056305120903408. URL: https://doi.org/10.1177/2056305120903408.

[124] J. Vincent. *Anyone can use this AI art generator — that's the risk*. Accessed: 13.08.2023. Sept. 2022. URL: https://www.theverge.com/2022/9/15/23340673/ai-image-generation-stable-diffusion-explained-ethics-copyright-data.

[125] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. "Image quality assessment: from error visibility to structural similarity." In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612. DOI: 10.1109/TIP.2003.819861.

[126] C. Warzel. "Faceapp shows we care about privacy but don't understand it." In: *New York Times* (2019).

[127] E. White. *Positive Implications Of Deepfake Technology In The Arts And Culture*. Accessed: 09.08.2023. Sept. 2021. URL: https://amt-lab.org/blog/2021/8/positive-implications-of-deepfake-technology-in-the-arts-and-culture.

[128] Wikipedia contributors. *Advanced Video Coding — Wikipedia, The Free Encyclopedia*. [Online; accessed 15-August-2023]. 2023. URL: https://en.wikipedia.org/w/index.php?title=Advanced_Video_Coding&oldid=1168205077.

[129] Wikipedia contributors. *Deepfake — Wikipedia, The Free Encyclopedia*. [Online; accessed 13-August-2023]. 2023. URL: https://en.wikipedia.org/w/index.php?title=Deepfake&oldid=1170192786.

[130] Wikipedia contributors. *FFmpeg — Wikipedia, The Free Encyclopedia*. [Online; accessed 15-August-2023]. 2023. URL: https://en.wikipedia.org/w/index.php?title=FFmpeg&oldid=1156808546.

[131] Wikipedia contributors. *Generative adversarial network — Wikipedia, The Free Encyclopedia*. [Online; accessed 13-August-2023]. 2023. URL: https://en.wikipedia.org/w/index.php?title=Generative_adversarial_network&oldid=1169846514.

[132] Wikipedia contributors. *Inpainting — Wikipedia, The Free Encyclopedia*. [Online; accessed 14-August-2023]. 2023. URL: https://en.wikipedia.org/w/index.php?title=Inpainting%5C&oldid=1164523541.

[133] Wikipedia contributors. *Latent space — Wikipedia, The Free Encyclopedia*. [Online; accessed 13-August-2023]. 2023. URL: https://en.wikipedia.org/w/index.php?title=Latent_space&oldid=1169655673.

[134] Wikipedia contributors. *Nvidia DGX — Wikipedia, The Free Encyclopedia*. [Online; accessed 15-August-2023]. 2023. URL: https://en.wikipedia.org/w/index.php?title=Nvidia_DGX&oldid=1168440769.

[135] Wikipedia contributors. *Stable Diffusion — Wikipedia, The Free Encyclopedia*. [Online; accessed 14-August-2023]. 2023. URL: https://en.wikipedia.org/w/index.php?title=Stable_Diffusion&oldid=1169859793.

[136] Wikipedia contributors. *UV mapping — Wikipedia, The Free Encyclopedia*. [Online; accessed 14-August-2023]. 2023. URL: https://en.wikipedia.org/w/index.php?title=UV_mapping&oldid=1169139226.

[137] S. Wirth. "Interface Experiments: FaceApp as Everyday AI." In: *Interface Critique* 4 (2023).

[138] C. H. Wu and F. D. la Torre. *Unifying Diffusion Models' Latent Space, with Applications to CycleDiffusion and Guidance*. 2022. arXiv: 2210.05559 [cs.CV].

[139] C.-Y. Wu, Q. Xu, and U. Neumann. *Synergy between 3DMM and 3D Landmarks for Accurate 3D Facial Geometry*. 2021. arXiv: 2110.09772 [cs.CV].

[140] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le. *Self-training with Noisy Student improves ImageNet classification*. 2020. arXiv: 1911.04252 [cs.LG].

[141] B. Yaylymov and D. T. D. Tran. *GitHub Repository - Deepfake Dataset*. Accessed: 14.08.2023. 2023. URL: https://github.com/yaylymov/deepfakes-dataset.

[142] Z. Yi, H. Zhang, P. Tan, and M. Gong. *DualGAN: Unsupervised Dual Learning for Image-to-Image Translation*. 2018. arXiv: 1704.02510 [cs.CV].

[143] L. Youngah, K.-T. Huang, R. Blom, R. Schriner, and C. Ciccarelli. "To Believe or Not to Believe: Framing Analysis of Content and Audience Response of Top 10 Deepfake Videos on YouTube." In: *Cyberpsychology, Behavior, and Social Networking* 24 (Feb. 2021). DOI: 10.1089/cyber.2020.0176.

[144] Zemana. *Deepware*. Accessed: 14.08.2023. URL: https://play.google.com/store/apps/details?id=com.zemana.deepware%5C&hl=gsw%5C&gl=US.

[145] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. *Self-Attention Generative Adversarial Networks*. 2019. arXiv: 1805.08318 [stat.ML].

[146] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. *StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks*. 2017. arXiv: 1612.03242 [cs.CV].

[147] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks." In: *IEEE Signal Processing Letters* 23.10 (2016), pp. 1499–1503. DOI: 10.1109/LSP.2016.2603342.

[148] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. *Random Erasing Data Augmentation*. 2017. arXiv: 1708.04896 [cs.CV].

[149] T. Zhou, W. Wang, Z. Liang, and J. Shen. "Face Forensics in the Wild." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2021, pp. 5778–5788.

[150] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. *Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks*. 2020. arXiv: 1703.10593 [cs.CV].

[151] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang. "Wilddeepfake: A challenging real-world dataset for deepfake detection." In: *Proceedings of the 28th ACM international conference on multimedia*. 2020, pp. 2382–2390.

[152] A. Zucconi. *Understanding the Technology Behind DeepFakes*. Accessed: 13.08.2023. Mar. 2018. URL: https://www.alanzucconi.com/2018/03/14/understanding-the-technology-behind-deepfakes/.