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

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, cybersecurity, 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. By analyzing the strengths and limitations of these techniques, this study sets the foundation for evaluating the effectiveness of publicly-available deepfake detection tools.

A robust methodology is employed to collect and analyze data on the available tools. The evaluation criteria include accuracy, efficiency, scalability, versatility, and user-friendliness. The selected deepfake detection tools, encompassing open-source projects, commercial offerings, and academic research projects, are assessed in detail to provide insights into their features, capabilities, and performance.

The findings of this study reveal the strengths and weaknesses of the evaluated deepfake detection tools. Comparative analysis sheds light on their distinctive characteristics and effectiveness in detecting deepfakes across different media types. Additionally, the study identifies gaps and challenges within the current landscape of deepfake detection, offering recommendations for future research, development, and policy-making.

The implications of this research extend to a wide range of domains, including media forensics, journalism, law enforcement, and online platforms. The ability to distinguish between genuine and manipulated content is crucial for safeguarding information integrity, maintaining trust, and combating disinformation campaigns. The insights provided by this thesis contribute to the ongoing efforts to develop effective deepfake detection mechanisms that keep pace with the evolving landscape of deepfake technology.

In conclusion, this thesis provides a comprehensive overview of publicly-available deepfake detection tools, offering an in-depth evaluation and comparison of their features and capabilities. The study highlights the urgent need for ongoing research

and development in the field of deepfake detection to counter the growing threat posed by synthetic media manipulation. By promoting a deeper understanding of the state of the art in deepfake detection, this research aims to contribute to the advancement of techniques and policies that can effectively mitigate the risks associated with deepfakes and uphold the integrity of digital media.

Contents

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 document seeks to provide an exhaustive review of the current state of the art in publicly-available deepfake detection tools.

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', is a technology that manipulates or fabricates audio-visual content to make it appear real, often indistinguishable from the original. An example of a generated and altered image can be seen in

Figure 1.1: Deepfake of Bill Hader impersonating Arnold Schwarzenegger. Screenshot from [6]

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 turned into a video created by using an nameless Reddit user

called 'deepfakes' in late 2017. This consumer started out to publish digitally altered pornographic motion pictures, realistically swapping the faces of actresses onto the bodies of porn stars. However, it wasn't lengthy earlier than the era was used out of doors of pornographic content.

An awesome instance that virtually established the electricity 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 [3], [12]. The video features a deepfake of Obama announcing matters he by no means clearly stated, with Peele providing the voiceover. This deepfake video, considered by tens of millions, efficaciously highlighted the capability misuse of this generation 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. 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]. 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 [10].

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 signs of digital artifacts and abnormal facial movements. But as detection tools become

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

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 comprehensive grasp 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 culminating in a synthesis of findings and forward-looking discussions. Below is a detailed outline of the thesis structure, which serves as a roadmap for navigating the document.

This initial section 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, 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 comprehensive survey of the existing body of knowledge related to deepfakes and their detection. The section begins with the history of deepfakes, tracing their evolution over time. It then delves into 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.

Section three, the research design and methods adopted for the study are outlined. The section provides detailed 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 section ensures that the research process is transparent and replicable.

Section four offers a comprehensive analysis of the selected deepfake detection tools. Each tool is explored in detail, discussing its working mechanisms, strengths, and potential limitations. This section also provides a comparative analysis of the tools, highlighting their relative strengths and weaknesses. Such a thorough examination is crucial to offer an in-depth understanding of the current landscape of publicly-available deepfake detection tools.

The fifth section 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 sectors and scenarios, thereby providing a broad view of the practical implications and challenges associated with deepfakes and their detection.

The sixth section presents the empirical findings from the evaluation of the selected tools. It provides a detailed report of how each tool performed across various tests, offering valuable insights into their effectiveness. This section serves as a vital point in the thesis, where collected data is introduced to support or challenge theoretical assertions.

The final section synthesizes the findings and discussions from the previous section and reflects on their contribution to the field. It provides a summary of the research, revisits the objectives, and discusses the extent to which they were achieved. 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.

In sum, the thesis follows a clear and logical structure that mirrors the research process, moving from the contextualization of the problem, through detailed analysis and evaluation, to the synthesis of findings and concluding reflections. This structure enables a thorough, systematic exploration of the state of the art of publicly-available deepfake detection tools, ensuring that the study is both comprehensive and focused.

1.3 Objectives of the Study

The primary objective of this thesis is to provide a comprehensive and in-depth exploration of the state of the art in publicly-available deepfake detection tools. This ambitious aim necessitates a multi-pronged approach, encompassing a wide array of secondary objectives that collectively serve to create a well-rounded examination of the topic. The identification and elaboration of these objectives provide a roadmap for the

study, with each one serving as a crucial stepping-stone toward the main goal.

The first objective is to trace the development of deepfake technology from its roots to its current state. This involves an in-depth exploration of the early techniques used in deepfake generation, the seminal developments that spurred its evolution, and the resulting modern methods capable of producing incredibly realistic and convincing deepfakes. Understanding the sophistication of the technology that we're attempting to counter is crucial, and can provide vital context for the subsequent investigation of detection tools.

While closely related to the first objective, the second objective delves deeper into the technical aspects of deepfake generation. The objective is to dissect and comprehend the underlying algorithms, techniques, and processes involved in creating deepfakes. This involves exploring machine learning and deep learning methods, such as autoencoders and Generative Adversarial Networks (GAN)s, that are fundamental to deepfake technology. This deep understanding can then be leveraged to better comprehend the mechanisms of deepfake detection tools.

At the heart of this thesis lies the primary investigative objective: the identification and detailed exploration of existing, publicly-available deepfake detection tools. This involves a comprehensive audit of these tools, an examination of their origins, the technology they employ, and their evolution in response to ever-improving deepfake generation techniques. This objective is crucial, as it provides the groundwork for the evaluation stage, providing us with a detailed understanding of what we're evaluating and why.

Having laid a thorough foundation with the previous objectives, the next goal is to objectively evaluate the performance of the identified deepfake detection tools. This assessment will be conducted using a wide array of deepfakes, evaluating the effectiveness, accuracy, and reliability of each tool across a spectrum of test cases. This rigorous evaluation process aims to determine how these tools fare against various types of deepfakes, offering insights into their strengths, weaknesses, and areas for potential improvement.

Given the potential for deepfakes to have significant societal impacts, a key objective of this study is to delve into the ethical, legal, and societal implications surrounding deepfakes and their detection. This includes exploring the potential risks deepfakes pose to information authenticity and privacy, as well as the ethical quandaries arising from the use of AI in deepfake detection. By illuminating these broader implications, the study aims to offer a more holistic view of the deepfake landscape.

The final objective of this study is to use the findings to propose concrete, actionable recommendations for future development in deepfake detection. These could range from technical enhancements for existing tools, the development of new, innovative detection methodologies, or even policy recommendations aimed at governing the use

and detection of deepfakes. By offering well-founded recommendations, this study aims to play a part in shaping the future direction of deepfake detection.

Collectively, these objectives provide a comprehensive framework for the study, allowing for a various exploration of the world of deepfakes and their detection. Each objective is not an end in itself but serves as a stepping stone towards the overall goal: to deepen our understanding of the state of the art in publicly-available deepfake detection tools and to contribute meaningfully to the ongoing efforts to mitigate the risks posed by deepfakes.

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 a thorough 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 fully comprehend 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 a comprehensive 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 comprehensive 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 representative sample of publicly-available deepfake detection tools, rather than an exhaustive 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 focused, in-depth, and accurate 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: Generative Adversarial Networks (GANs) and autoencoders.

GANs introduced by Goodfellow et al. [11], nvolve two competing neural networks: a generator and a discriminator. The generator produces fake samples and the discriminator distinguishes between the real and fake samples. 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 [2].

Autoencoders, on the other hand, are a type of neural network used for learning efficient encodings or representations of input data [13]. In the context of deepfakes, autoencoders are used to learn the compressed representation of faces and are then able to regenerate them based on the learned model. 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.

Recent developments have seen the rise of Variational Autoencoders (VAE) and their use in deepfake generation [16]. Unlike traditional autoencoders, VAEs introduce a probabilistic spin to the encoding and decoding processes. 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.

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

Known for offering greater functionality and control over the deepfake creation process, DeepFaceLab[1] 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 comprehensive suite of tools, combined with its high-quality results, make it a popular choice among deepfake creators.

Figure 2.1: Deepfake of Ivanka Trump impersonating Emma Watson. Screenshot from our own generated deepfake video with DeepFaceLab.

2.2.2 FaceSwap

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[2] 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.

[1] `https://github.com/iperov/DeepFaceLab`
[2] `https://faceswap.dev`

Figure 2.2: Deepfake of Emma Stone impersonating Scarlett Johansson using Faceswaps's Phaze-A model. Screenshot from [7].

2.2.3 Stable Diffuison

Stable Diffusion[3] models have also emerged as a powerful technique for generating deepfakes. They utilize a diffusion process to generate realistic synthetic images from a simple Gaussian noise, achieving impressive results in the generation of deepfake images [25]. 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 with. This makes them particularly well-suited for tasks like deepfake generation, where capturing the detailed, multi-modal distribution of human faces is essential.

Figure 2.3: Deepfake of Justin Trudeau created with Stable Diffusion. Screenshot from our own generated deepfake dataset.

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

2.2.4 Neural Textures

Introduced by Thies et al. [23], 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 2.4, can be trained to synthesize person-specific details, resulting in high-quality face swaps or manipulation of facial expressions in videos.

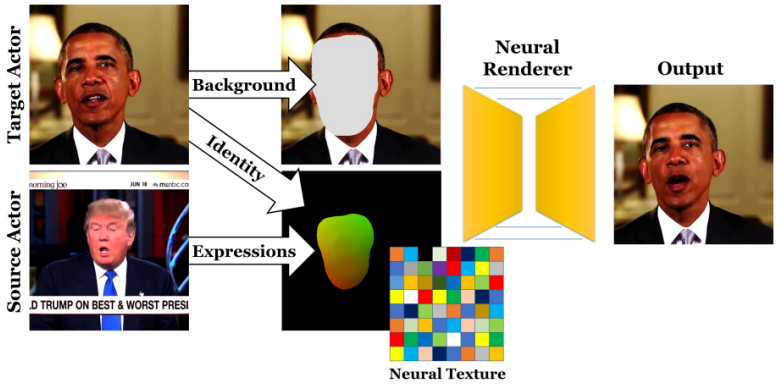

Figure 2.4: The reenactment synthesis process uses expression transfer to generate a UV map of the target actor that reflects the source actor's expression. This map, along with a background image, is processed by a neural renderer to create the final reenactment. Expression alteration is achieved by training a unique neural texture and renderer for the target actor, resulting in a manipulated video as shown in the image from [23].

2.2.5 FaceApp

While not a deepfake tool in the traditional sense, FaceApp[4] 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. 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 [20]. FaceApp has been both praised for its technological achievements and criticized for its potential privacy and consent issues, reflecting the wider debates surrounding the

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

ethical implications of deepfake technology.

Figure 2.5: Deepfake created with FaceApp.

2.3 Ethical and Legal Concerns

The rise of deepfakes has brought with it a host of ethical and legal concerns. At the forefront is the issue of consent, as deepfakes often involve the use of a person's likeness without their permission. This has been particularly prevalent in the creation of deepfake pornography, leading to significant harm and distress for the individuals involved [4].

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 [14], [18]. The realistic nature of deepfakes makes it difficult for the average viewer to discern truth from falsehood, further exacerbating these risks.

In the realm of journalism, the rise of deepfakes presents both a significant challenge and an ethical dilemma. Journalists must not only navigate the increasingly complex task of verifying the authenticity of digital content but also grapple with the ethical implications of using AI-generated content in their reporting. Misuse of deepfakes could lead to the spread of false information, severely undermining public trust in the media [24].

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 [4].

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 executive or other authority figure. These scams could lead to significant financial losses or damage to a company's reputation.

Lastly, in the realm of 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 accurate deepfake detection methods [22].

2.4 Existing Countermeasures and Detection Methods

The advent and proliferation of deepfakes necessitate effective countermeasures and detection methods to ensure information integrity and maintain public trust in digital media. Detection methods have evolved in response to the sophistication and complexity of deepfake generation techniques. Several of these methods employ machine learning and more specifically deep learning strategies, leveraging the same kind of technology used to create deepfakes, to counter them.

The general principle 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 statistical distribution of the pixel values, or unnatural physical characteristics such as inconsistent lighting or improper blinking patterns [1].

One approach to deepfake detection 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. [9], exploit the fact that deepfake generation algorithms usually operate in the spatial domain and thus might introduce specific anomalies in the frequency domain.

Another widely used approach is the Convolutional Neural Network (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 [21].

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 [5].

Recent advancements have led to the development of deepfake detection techniques that analyze physiological signals. For instance, Li et al. [19] 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 comprehensive, encompassing the nuances of accessibility, ease of use, limitations, variety in detection methods, and documentation. A detailed description of the selection criterias is given in Table 3.1.

3.2 Evaluation Metrics

After careful selection of the tools based on the prescribed criteria, it became imperative 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 employed. These metrics serve as a guiding light, illuminating the capabilities of each tool in terms of detecting deepfakes. The evaluation metrics employed 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. In Figure 3.1 a fake sample image of each dataset is provided.

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 an integrated development environment (IDE) and specialized packages? Additionally, any constraints, such as file size limits or video duration caps, play a role in its overall user-friendliness.
Accessibility	Only publicly accessible tools were taken into account, promoting the accessibility of deepfake detection and ensuring that a wide spectrum of users, from the general public to specialists, can utilize the tools. The cost factor is another crucial aspect of accessibility; tools that are freely available or open source often garner a larger user base compared to proprietary or paid solutions. Whether the tools are available through a simple browser interface or require local installation can greatly influence accessibility. Additionally, while some advanced tools might demand powerful GPU setups, the most accessible ones should be usable on standard hardware configurations or offer cloud-based solutions, like Google Colab, to bypass local hardware limitations.
Support and Documentation	Robust documentation and active support, be it community or developer-driven, are crucial. Comprehensive support ensures that users can fully utilize tool features, troubleshoot issues, and gain deeper insights into the tool's workings.
Dataset choice	The datasets a tool is compatible with or recommends can reflect its versatility and potential applications. Tools that can adapt to various datasets or come with robust recommended datasets are in a favorable position to tackle countless deepfake challenges.

Table 3.2: Evaluation Metrics

Evaluation Metrics	**Description**
Processing time and scalability	Measures the time taken by the tool to detect deepfakes in a given input. It also evaluates how well the tool performs when the size and the number of the input data increases.
Interpretability	Assesses how understandable and transparent the results or outputs of the tool are. It's crucial for users to comprehend why certain detections are made.
Detection Accuracy	The proportion of true results (both true positives and true negatives) in the total dataset. It provides a comprehensive measure of the tool's ability to correctly identify both genuine and deepfake content.
Precision	The proportion of true positive results in the total predicted positives (both true and false positives). It measures the tool's capability to avoid false positives, ennsuring that genuine content isn't mistakenly flagged.
Recall	Measures the tool's ability to correctly identify actual deepfakes out of all genuine deepfakes presented. It calculates the number of actual true positives the tool identifies.
F1-Score	The harmonic mean of Precision and Recall. It provides a balance between the two when there's an uneven class distribution.

3.3.1 FaceForensics++

FaceForensics++ [1] serves as a comprehensive 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. Originally launched in 2018, the dataset received significant updates in 2019, making it richer and more diverse. As highlighted by the creators in their 2019 paper [22], the videos in the dataset were modified using a mix of techniques. This includes two graphics-driven methods, namely Face2Face and FaceSwap, as well as two methods rooted in machine learning: DeepFakes and NeuralTextures. 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.

3.3.2 Deepfake Detection Challenge Dataset

The Deepfake Detection Challenge (DFDC) dataset emerged from a collaborative initiative hosted on Kaggle [15], 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 dollars. 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 [8]. This variety makes it a great tool to test and improve deepfake detection methods.

3.3.3 Face Forensics in the Wild

The Face Foresics in the Wild (FFIW) [2] dataset offers 10,000 high-quality manipulated videos. 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 [26].

3.3.4 OpenForensics

The OpenForensics [3] dataset is tailored for detecting and segmenting multi-face forgeries. Its version 1.0.0 houses more than 115,000 real-world images, capturing a total

[1] https://github.com/ondyari/FaceForensics
[2] https://github.com/tfzhou/FFIW
[3] https://github.com/ltnghia/openforensics

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 [17].

(a) FaceForensics++ (b) DFDC

(c) FFIW (d) OpenForensics

Figure 3.1: Sample deepfake images taken from FaceForensics++ [22], DFDC [8], FFIW [26] and OpenForensics [17]

4 Analysis of Publicly-Available Deepfake Tools

For a thorough analysis in this study, a variety of tools were picked. Their selection was not arbitrary; instead, it was based on the clear 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.

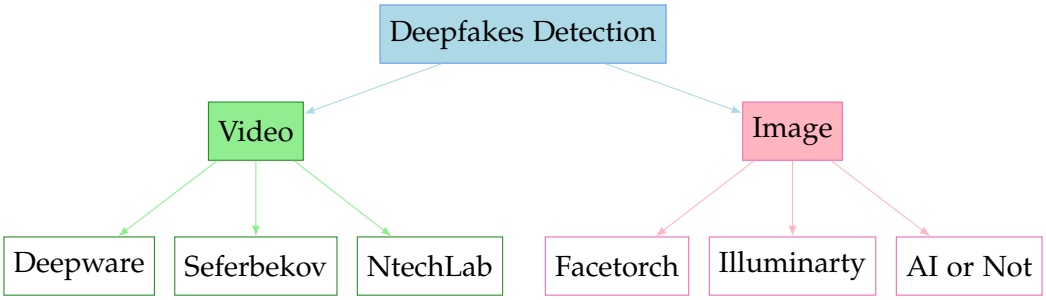

Figure 4.1: Categorization of deepfake detection tools

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 comprehensive analysis.

4.1 Deepware

4.2 Seferbekov

4.3 NtechLab

4.4 Facetorch

4.5 Illuminarty

4.6 AI or Not

4.7 Comparative Analysis

5 Case Studies

5.1 Entertainment and Art

5.2 Politics and Media

5.3 Cybersecurity and Privacy

6 Results

6.1 Dataset Augmentations

6.2 Frequency Analysis

6.3 Final Results

7 Conclusion

7.1 Discussion and Recommendations

7.2 Summary of Findings

7.3 Future Researcher Directions

8 Test

8.1 Section

Acronyms must be added in `main.tex` and are referenced using macros. The first occurrence is automatically replaced with the long version of the acronym, while all subsequent usages use the abbreviation.

E.g. `\ac{TUM}`, `\ac{TUM}` ⇒ Technical University of Munich (TUM), TUM

For more details, see the documentation of the `acronym` package[1].

8.1.1 Subsection

See Table 8.1, Figure 8.1, Figure 8.2, Figure 8.3.

Table 8.1: An example for a simple table.

A	B	C	D
1	2	1	2
2	3	2	3

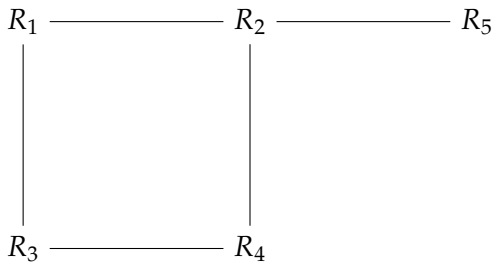

Figure 8.1: An example for a simple drawing.

[1] `https://ctan.org/pkg/acronym`

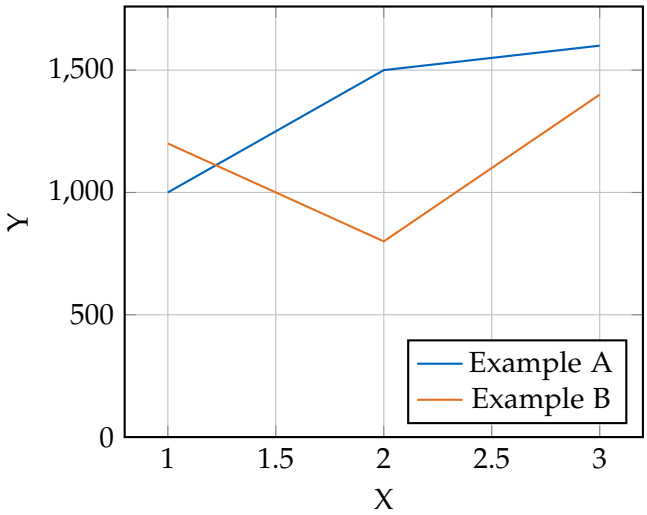

Figure 8.2: An example for a simple plot.

```
SELECT * FROM tbl WHERE tbl.str = "str"
```

Figure 8.3: An example for a source code listing.

Abbreviations

TUM Technical University of Munich

AI Artificial Intelligence

GAN Generative Adversarial Networks

VAE Variational Autoencoders

CNN Convolutional Neural Network

IDE Integrated Development Environment

DFDC Deepfake Detection Challenge

FFIW Face Foresics in the Wild

List of Figures

List of Tables

Bibliography

[1] 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.

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

[3] 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.

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

[5] 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].

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

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

[8] 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].

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

[10] 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/.

[11] 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].

[12] 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.

[13] 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.

[14] 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.

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

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

[17] 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.

[18] 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/.

[19] 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].

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

[21] 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].

[22] 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.

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

[24] 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.

[25] 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].

[26] 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.