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

# 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 was a video made by an anonymous Reddit user named '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], [18]. 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 [12].

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. [17], 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 [19]. 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 [22]. 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 [8].

### 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 [35]. 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. [33], 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.

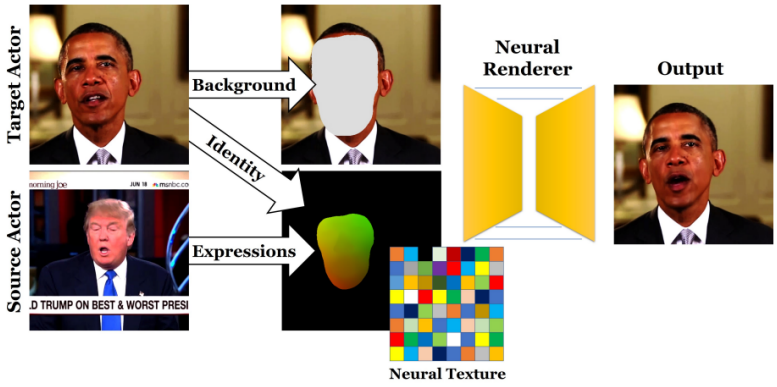



Figure 2.4: The reenactment synthesis process uses expression transfer to generate a UV map of the target actor that reflects the source actor's expression. This map, along with a background image, is processed by a neural renderer to create the final reenactment. Expression alteration is achieved by training a unique neural texture and renderer for the target actor, resulting in a manipulated video as shown in the image from [33].

### 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 [28]. 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.

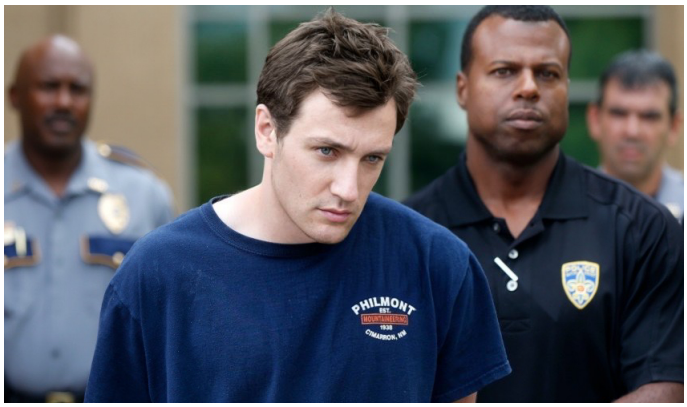



Figure 2.5: Deepfake created with FaceApp.

## 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 [20], [25]. The realistic nature of deepfakes makes it difficult for the average viewer to discern truth from falsehood, further exacerbating these risks.

In journalism, the rise of deepfakes presents both a significant challenge and an ethical dilemma. Journalists must not only navigate 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, deeply shaking people's faith in the media [34].

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 [30].

## 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. [11], 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 [29].

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. [26] 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 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 [30], 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 [21], 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 [10]. 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 [38].

### 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

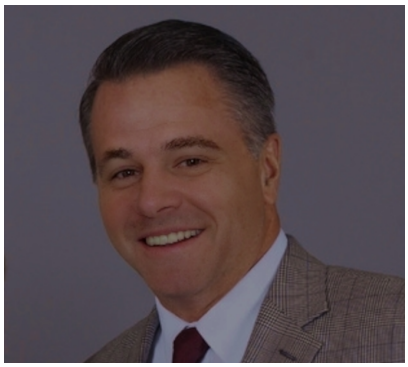
[1] `https://github.com/ondyari/FaceForensics`
[2] `https://github.com/tfzhou/FFIW`
[3] `https://github.com/ltnghia/openforensics`

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 [24].

(a) FaceForensics++

(b) DFDC

(c) FFIW



(d) OpenForensics

Figure 3.1: Sample deepfake images taken from FaceForensics++ [30], DFDC [10], FFIW [38] and OpenForensics [24].

# 4 Analysis of Publicly-Available Deepfake Detection 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.

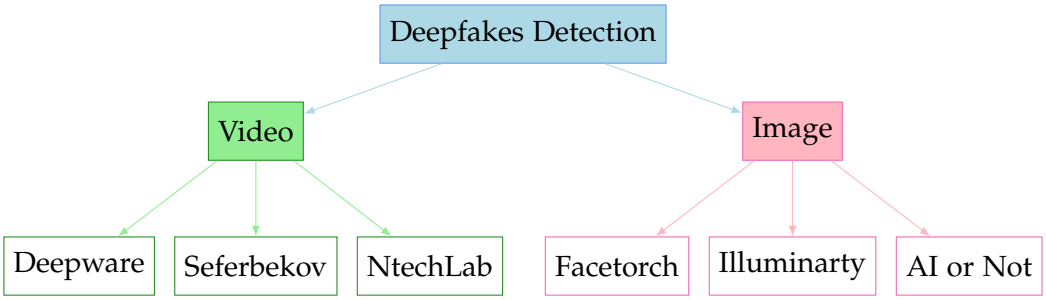



Figure 4.1: Categorization of deepfake detection tools

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.

For a thorough 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, utilizing the DeepFaceLab and FaceSwap tools. The remaining 30 were sourced from the established FaceForensics++ (Subsection 3.3.1) database, wherein 10 were identified as deepfakes, while 20 were authentic.

For the image detection analysis, we employed a total of 123 pictures. Out of these, 103 were deepfakes produced using FaceApp (Subsection 2.2.5) and Stable Diffusion (Subsection 2.2.3), while the remaining 20 were authentic images from our proprietary dataset.

## 4.1 Deepware

Deepware[1] is a tool made to tackle a big problem: the rise of fake videos or 'deepfakes'. The people behind Deepware saw this challenge early on. Their parent company, Zemana[2], 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.

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.

Using Deepware is straightforward. There's a user-friendly website where deepfakes can be easily uploaded for assessment. No advanced hardware requirements are imposed on the 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 comprehensive detection, Deepware employs 4 different deepfake detection models.

A notable feature integrated into Deepware's platform allows for expert reviews. If inaccuracies in the deepfake detection process are suspected, a specialized review can be requested, underscoring the team's commitment to accuracy and continuous improvement.

In the spirit of community collaboration, parts of Deepware's detection mechanism have been open-sourced. By making their findings and tools accessible, a collaborative approach to enhancing deepfake detection is actively encouraged.

The outcomes of the metrics 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 |

**Detection Accuracy** is the overall correct classification of the tool, considering both True Positives (TP) (deepfakes correctly identified) and True Negatives (TN) (genuine videos correctly identified). It is calculated as follows:

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

[1] `https://deepware.ai/`

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

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% |

**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[3] deepfake detection tool emerged as a winner in the DFDC challenge, securing a whopping prize of $500,000 [21]. 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 [32]. 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 [31], a huge

[3] `https://github.com/selimsef/dfdc_deepfake_challenge`

database of images, and a method called 'noisy student'. This 'noisy student' technique, as detailed in a research paper [36], allows the tool to learn better and improve its accuracy.

For each video it studies, 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 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 Graphics Processing Unit (GPU)s 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 evaluation metrics are provided in Section 4.1.

In terms of interpretability, Seferbekov's tool delivers comprehensible 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[4] software made waves by clinching third place in the DFDC Challenge, walking away with a cool $100,000 prize [21]. 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 utilized superior hardware, the decision was made to employ 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 layers on how this tool works: the heart of the software is based on a trio of EfficientNet-B7 models [32]. These models use 'Noisy Student' [36] 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 [7] and Random Erasing [37]. They even had a clever way of adjusting video compression on-the-spot, which involves playing around with short video clips.

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 results from the metrics tested with NtechLab can be seen in Table 4.5 and Table 4.6.

[4] `https://github.com/NTech-Lab/deepfake-detection-challenge`

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[5], crafted by Tomáš Gajarský, stands out as a cutting-edge tool designed for image forgery detection. It is a a straightforward Python tool built on PyTorch[6], a popular programming framework. It's designed to identify faces and explore the detailed aspects of facial features using specific algorithms known as neural networks. The aim is to bring together the finest pre-existing models, enhance their speed with TorchScript[7], and present an all-in-one solution for face analysis. The features it offers include:

- Face Detection
- Facial Representation Learning
- Face Verification
- Facial Expression Recognition
- Deepfake Detection
- 3D Face Alignment

[5] https://github.com/tomas-gajarsky/facetorch
[6] https://pytorch.org/
[7] https://pytorch.org/docs/stable/jit.html

Notably, this tool comes with a user manual [16], Application Programming Interface (API) instructions [13], and an instance on Hugging Face [15], 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 [14]. One of its most appealing aspects is its open-source nature, with updates being consistently rolled out up until March 2023.

The results from the metrics tested with NtechLab can be seen 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 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[8] is an online platform where users can easily upload images to detect deepfakes and AI generated images. 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 utilize AI and Deepfake images and texts. There is also a user-friendly "Terms of Use"

[8] `https://app.illuminarty.ai/`

section that explains the tool's purpose and usage instructions. Additionally, if users need suport, they can reach out the Illuminarty community through Discord or Patreon.

The results from the metrics tested with Illuminarty is provided 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[9] is the idea of Andrey Doronichev, formerly Director of Product at Google. In 2022 he founded Optic[10], a startup dedicated to identifying the authenticity of images, videos and voices. Optic offers three main products:

**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 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[11], DALL-E[12] and GAN [9].

[9] https://www.aiornot.com/
[10] https://www.optic.xyz/
[11] https://www.midjourney.com/
[12] https://openai.com/dall-e-2

It also offers Chrome extension and a Telegram bot[13], 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.

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.

Here are the results from the tested metrics with AI or Not:

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% |

[13] https://t.me/AI_or_not_bot

# 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 [27].

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 [23].

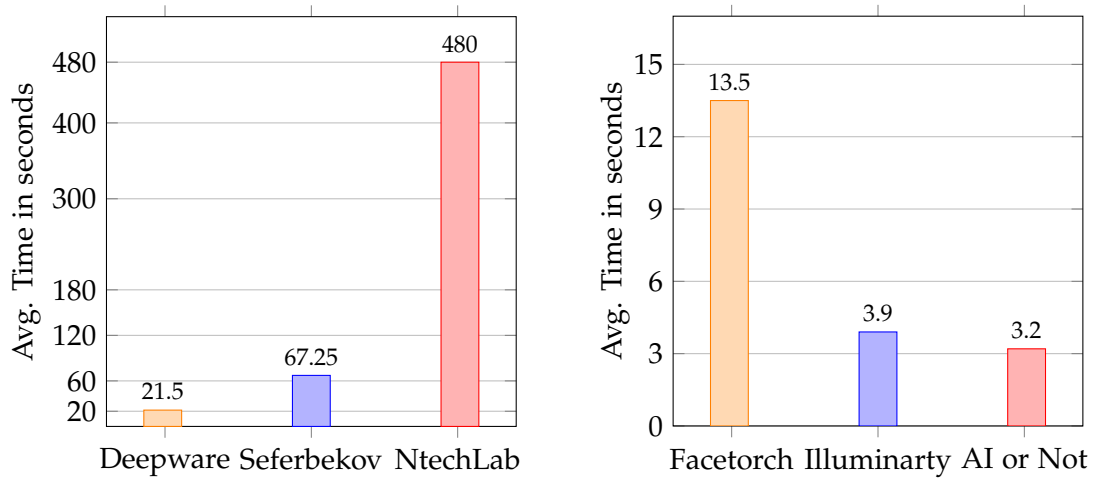
## 5.2 Politics and Media

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

# 6 Results

In this chapter, the results of the in Chapter 4 computed experinments are presented. 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 employed.

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.

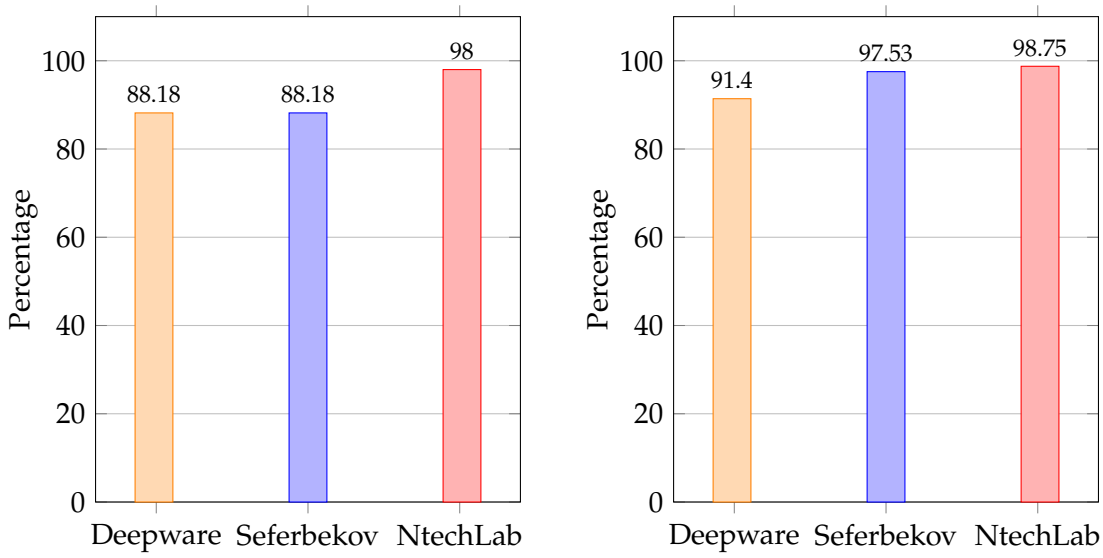
A comparison of assessed metrics for video detection tools is displayed in Figure 6.2. The comparison of which tool performed better is interesting. As we know, accuracy and precision are defined as follows: Accuracy measures the fraction of all instances that are correctly identified and precision measures 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

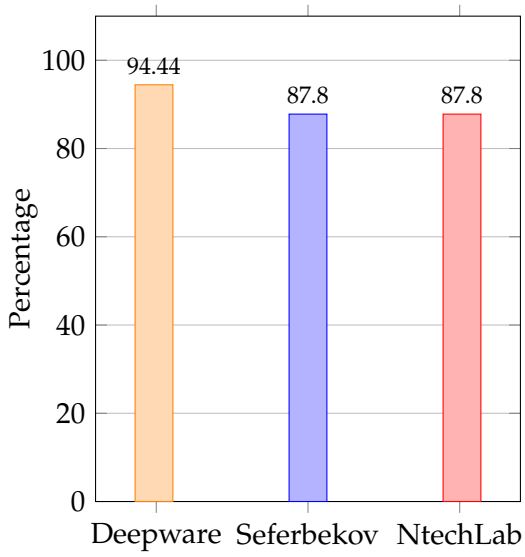
To conclude the achieved results, it's noticable that some tools might have a high accuracy rate but perform poorly in terms of recall and F1-Score. Contrarily, certain tools might demonstrate high precision but low recall, which might indicate that the tool misses some actual positives. By comparing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



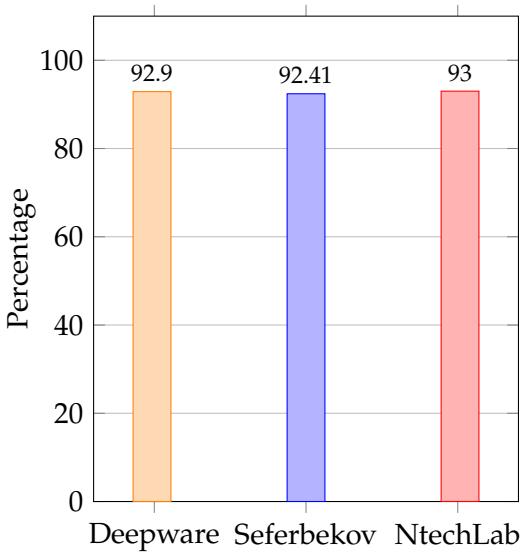

(a) Detection accuracy




(b) Precision

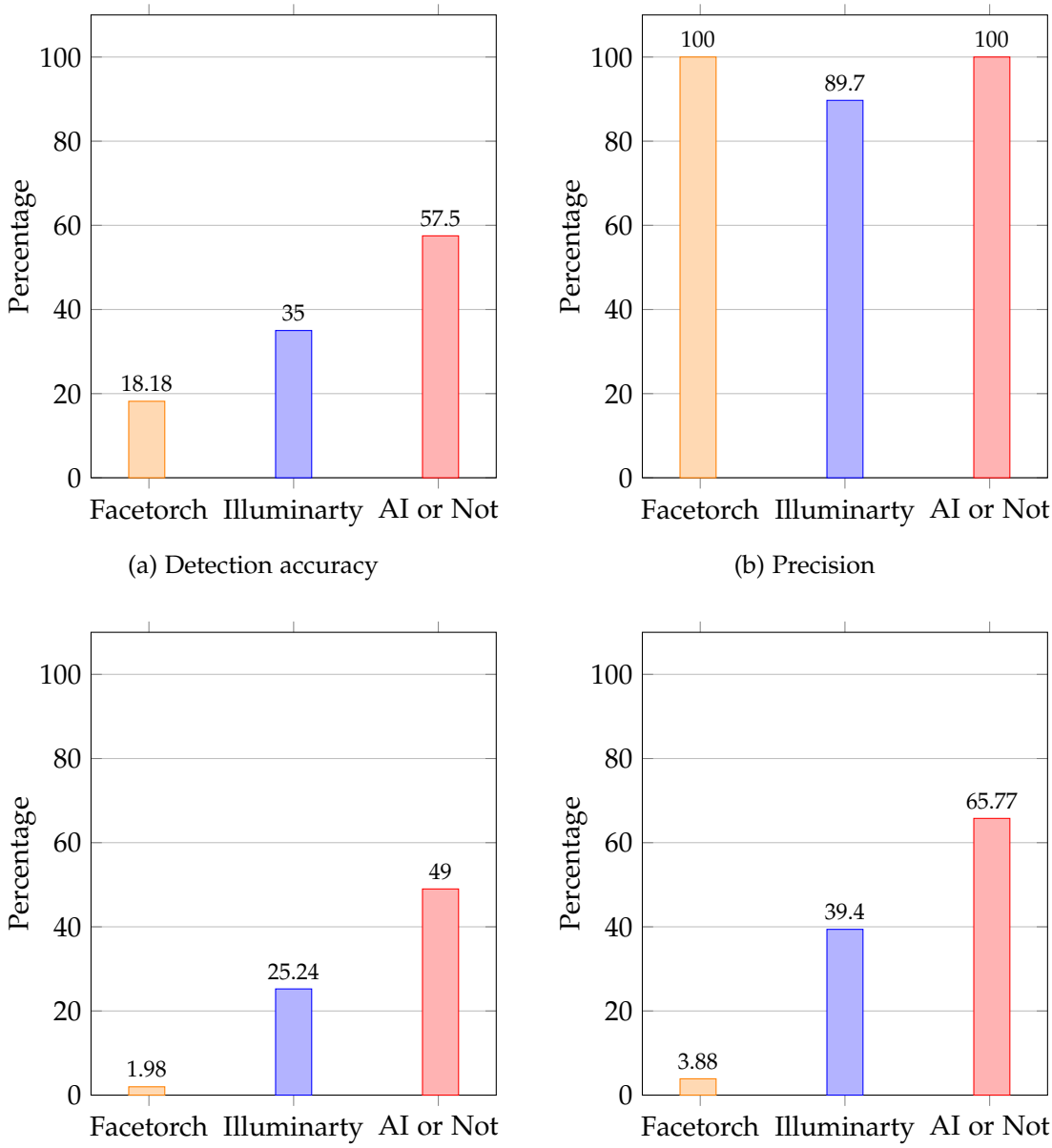



(c) Recall




(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

# 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

**VAE** Variational Autoencoders

**CNN** Convolutional Neural Network

**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

# 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] 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].

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

[9] 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/.

[10] 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].

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

[12] 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/.

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

[14] 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.

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

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

[17] 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].

[18] 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.

[19] 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.

[20] 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.

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

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

[23] 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/.

[24] 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.

[25] 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/.

[26] 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].

[27] 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/.

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

[29] 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].

[30] 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.

[31] 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.

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

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

[34] 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.

[35] 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].

[36] 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].

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

[38] 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.