Projektarbeit

Deepfakes und Social Engineering

vorgelegt von

Julian Faigle (Matrikelnummer: 86292)
Studiengang ITS

Max Ernstschneider (Matrikelnummer: 86464)
Studiengang AIT

Semester 6

Hochschule Aalen

Hochschule für Technik und Wirtschaft

Betreut durch Prof. Roland Hellman

15.08.2024

Erklärung

Wir versichern, dass wir die Ausarbeitung mit dem Thema „Deepfakes und Social Engineering“ selbstständig verfasst haben und keine anderen Quellen und Hilfsmittel als die angegebenen benutzt haben. Die Stellen, die anderen Werken dem Wortlaut oder dem Sinn nach entnommen wurden, sind in jedem einzelnen Fall unter Angabe der Quelle als Entlehnung (Zitat) kenntlich gemacht worden. Das Gleiche gilt für beigefügte Skizzen und Darstellungen.

Aalen, den 18. Juli 2024

Ort, Datum

Julian Faigle

Autor

Max Ernstschneider

Autor

Inhaltsverzeichnis

1 Grundlagen **1**

1.1 Einführung in Deepfakes . . . 1
1.1.1 Definition . . . 1
1.1.2 Hintergrund . . . 1
Geschichte . . . 1
1.1.3 Technische Grundlagen . . . 3
1.1.4 Arten von Deepfakes . . . 3
Video Deepfakes . . . 3
Face Swapping . . . 3
Face Morphing . . . 3
Reenactment . . . 4
Full body puppetry . . . 4
Audio Deepfakes . . . 4
Voice-swapping . . . 5
Text to Speech . . . 5
Foto Deepfakes . . . 5
Face and body-swapping . . . 5
Kombination aus Audio und Video Deepfake . . . 5
1.1.5 Anwendungsgebiete . . . 6
Positive Anwendungsgebiete . . . 6
Kunst- und Filmbranche . . . 6
Negative Anwendungsgebiete . . . 6
Politik und Regierung . . . 6
Wirtschaft . . . 6
Erstellung künstlicher Indentitäten . . . 6
Mobbing . . . 6
Pornographie . . . 6
1.1.6 Ethik . . . 6
1.2 Social Engineering . . . 6
1.2.1 Verschiedene Typen von Social Engineering . . . 7
1.2.2 Überblick über gängige Angriffe . . . 8
1.2.3 Gegenmaßnahmen gegen Social Engineering Angriffe . . . 9
1.3 Bekannte Social Engineering Angriffe . . . 10

2 Deepfake Varianten **12**
2.1 Video-Deepfakes . . . 12
2.1.1 Veraltete Methode: 3D Modellierung . . . 12
2.1.2 Generative Adversarial Networks (GANs (Generative Adversarial Networkss)) . . . 13
2.1.3 FSGAN (Face Swapping GAN) und FSGANv2 . . . 13
2.1.4 Gemeinsamkeiten und Unterschiede . . . 14
2.2 Face Swapping vs. Reenactment . . . 14
2.2.1 Face Swapping . . . 15
2.2.2 Reenactment . . . 15
2.3 Detection Methods für Video-Deepfakes . . . 16

3 Erstellung von Deepfake Videos **18**
3.1 DeepFaceLab . . . 18
3.1.1 Motivation . . . 18
3.1.2 Fähigkeiten . . . 18
3.1.3 Workflow . . . 18
Extraction . . . 19
Training . . . 19
Konvertierung . . . 20
3.2 Praxisbeispiel . . . 20
3.2.1 Laborumgebung . . . 20
3.2.2 Extraction . . . 21

Akronyme

CLI	Command Line Interface
CNN	Convolutional Neural Network
DBIR	Data Breach Investigations Report
DFL	DeepFaceLab
FSGAN	Face Swapping GAN
GAN	Generative Adversarial Networks
IDS	Intrusion Detection System
JIT	Just-in-Time
LRCN	Long-Term Recurrent Convolutional Network
LSTM	Long Short-Term Memory
RDJ	Robert Downey Jr.

Glossar

Enkeltrick	Ein betrügerisches Vorgehen, bei dem sich Trickbetrüger über das Telefon, neuerdings auch über Kontaktplattformen und Messengerdienste, meist gegenüber älteren und/oder hilflosen Personen, als deren nahe Verwandte ausgeben, um unter Vorspiegelung falscher Tatsachen an deren Bargeld oder Wertgegenstände zu gelangen
Motion Tracking	Motion Tracking ist eine Technik, die verwendet wird, um die Bewegung von Objekten oder Personen in einem Video oder einer animierten Szene zu verfolgen und zu verfolgen. Dies kann in 2D oder 3D erfolgen.
Threat Intelligence	Threat Intelligence sind Daten, die gesammelt, verarbeitet und analysiert werden, um die Motive, Ziele und das Angriffsverhalten eines Bedrohungsakteurs zu verstehen. Durch Threat Intelligence können schnellere, fundiertere und datenbasierte Sicherheitsentscheidungen getroffen werden. Zudem ermöglicht es, das Verhalten im Kampf gegen Bedrohungsakteure von reaktiv zu proaktiv zu ändern.[1]

1. Grundlagen

1.1 Einführung in Deepfakes

1.1.1 Definition

Der Begriff Deepfake setzt sich aus den englischen Begriffen Deep Learning und Fake zusammen. Hierbei steht Deep Learning für eine Methode des maschinellen Lernens und Fake für eine Fälschung.
"Bei Deepfakes handelt es sich um einen Teilbereich synthetischer audiovisueller Medien: die Manipulation oder auch synthetische Erzeugung von Abbildungen, Videos und/oder Audiospuren menschlicher Gesichter, Körper oder Stimmen, zumeist mithilfe von KI."[2] [2]
Deepfakes werden mit Hilfe von künstlicher Intelligenz und Deep Learning Technologien erstellt, um Personen realistische Handlungen ausführen oder Worte sagen zu lassen, in Form von Video, Bild oder Audio. Es handelt sich hierbei um gefälschte Darstellungen, die möglichst realitätsnah dargestellt werden.[3]

1.1.2 Hintergrund

Deepfake ist eine Manipulationstechnik, die es Benutzern ermöglicht, das Gesicht einer Person mit einer anderen Person auszutauschen. Eine optimale Manipulation wird durch Verwendung mehreren Hunderten oder Tausenden Fotos der Zielperson erreicht. Das führt dazu, dass oft prominete Personen als Zielperson gewählt werden, da von ihnen viele Bilder im Internet existieren.
Bild- und Videomanipulationstechnologien bauen auf Techniken aus dem Bereich der künstlichen Intelligenz auf, welcher das Ziel verfolgt, menschliche Denkprozesse und Verhaltensweisen zu verstehen. Da maschinelles Lernen einem System ermöglicht aus Daten zu lernen, ist diese Technik wichtig für das erstellen von Deepfakes.
Deepfakes sind aus zwei Gründen beliebt: erstens wegen der Fähigkeit aus Daten wie Fotos und Videos, realistische Ergebnisse erzeugen zu können und zweitens die Verfügbarkeit der Technik, da diese für jeden leicht zu erreichen und durchzuführen ist. Es gibt Apps, welche die Schritte des Deepfakes-Algorithmus erklärt und so Personen mit wenig Kenntnissen über maschinelles Lernen oder Programmierung die Möglichkeit bietet ein Deepfake Bild oder Video zu erstellen.
Das führt zu einem Problem der heutigen Gesellschaft, da Deepfakes hauptsächlich aus Rache, Erpressung einer Person oder Verbreitung von Fake News einer höheren Person (bspw. eines Politikers) ausgenutzt werden.[4]

Geschichte

Das Manipulieren von Bildern wurde nicht erst in den letzten Jahren bekannt. Denn auch schon früher wurden Bilder zum Beispiel von Hitler, Stalin, oder Breschnew mani-

puliert, um so die Geschichte zu ihren Gunsten verändern zu können. Damals erforderte es allerdings deutlich mehr Zeit und kompliziertere Techniken während der Fotoentwicklung in der Dunkelkammer, um ein Bild zu verfälschen. Doch durch die schnelle Entwicklung der Technologien wurde der Prozess ein Bild zu manipulieren zunehmend schneller. Anfangs begannen ausschließlich Forscher der 1990er Jahre die Entwicklung der Deepfake-Technologie zu übernehmen, diese wurde jedoch später von Amateuren in den Online-Communities unterstützt. Die Akademiker Christoph Bregler, Michele Covell und Malcolm Slaney entwickelten 1997 ein Programm, welches vorhandenes Videomaterial einer sprechenden Person anpassen konnte, dass diese Person die Wörter von einer anderen Audiospur nachahmte. Das Programm baut auf einer älteren Technologie auf, welches bereits Gesichter interpretieren, Audio aus Texten synthetisieren und Lippen im 3D-Raum modellieren konnte. Jedoch war dieses entwickelte Programm von den drei Akademikern das erste, welches alle Komponenten zusammenfügen und überzeugend animieren konnte. So war es möglich eine neue Gesichtsanimation aus einer Audioausgabe zusammenstellen zu können.
Zu Beginn der 2000er Jahre wurde die Entwicklung der Gesichtserkennung mit dem Computer immer weiter vorangetrieben, sodass es zu großen Verbesserungen der Technologie wie Motion Trackings kam, welche die heutigen Deepfakes so überzeugend machen.
In den Jahren 2016 und 2017 gab es zwei Projekt Veröffentlichungen. Einmal das Face2Face-Projekt der Technischen Universität München und einmal das Synthesizing Obama-Projekt der University of Washington.
Das Face2Face Projekt versucht Echtzeitanimationen zu erstellen, indem es den Mundbereich des Zielvideos durch einen Schauspieler ersetzt, während das Synthesizing Obama-Projekt sich damit beschäftigte Videomaterial des ehemaligen Präsidenten Barack Obama zu modifizieren.[5]
Im Jahr 2017 wurde das gefälschte Video des ehemaligen US-Präsidenten Barack Obama veröffentlicht und soll als Warnung der Technologie und deren potenziellen Auswirkungen gelten. Ende 2017 veröffentlichte ein Nutzer auf einer Webseite names Reddit pornografische Inhalte und behauptete, dass diese zu bekannten Personen wie zum Beispiel Taylor Swift oder Scarlett Johansson gehören. Auch wenn diese Bilder und Videos schnell wieder gelöscht wurden, erregte diese auf Deep Learning basierende Gesichtsersatztechnik die Aufmerksamkeit der Medien und verbreitete sich in vielen Internetforen. Alle Inhalte, die mit der Deepfake Technik zu tun hatten, wurden am 7. Februar 2018 auf fast allen Internetforen entfernt und verboten. Trotz des Verbots hat sich die Technik dennoch weiterhin durchgesetzt und wurde weltweit verbreitet. Bei der Person, die die Deepfake-Technik entwickelt hat, soll es sich um einen Software-Ingenieur handeln, der ein Entwicklungs-Kit herausbrachte, mit dem es einem Benutzer selbst ermöglicht, eigene manipulierte Bilder oder Videos zu erstellen. Durch die Hilfe von Open Source Tools und Funktionen von großen Softwareunternehmen wie NVidia und Google wurde die Deepfake-Technike entwickelt. Was bedeutet, dass für die Entwicklung technisches Wissen und Verständnis erforderlich sind, jedoch der Großteil der Software schon zuvor in der Öffentlichkeit zur Verfügung stand. Als klar wurde, dass selbst eine Person ohne viel Wissen in dem Gebiet, beliebig viele visuelle Medien manupulieren kann, wurde die

Bedrohung der Deepfake-Technik ernst und das US-Verteidigungsministerium stellte sich ein. Auch im Jahr 2018 wurde ein Deepfake Video von damaligen Präsidenten Donald Trump in den Medien hochgeladen, in dem die Belgier aufgefordert wurden, aus dem Pariser Klimaschutzabkommen auszusteigen.
Durch solche Veröffentlichungen der Deepfake Videos zeigte sich, dass die Technologie sich schnell weiterentwickelt und in der Lage ist einen großen Teil der Öffentlichkeit in die Irre führen zu können.[4]

1.1.3 Technische Grundlagen

1.1.4 Arten von Deepfakes

Deepfakes können in drei Hauptarten unterteilt werden: Video Deepfakes, Audio Deepfakes und Foto Deepfakes. Diese drei Arten lassen sich zusätzlich auch noch miteinander kombinieren.[5]

Video Deepfakes

Bei Video Deepfakes wird zusätzlich zwischen 3 Arten der Manipulation unterschieden. Auf welche Art der Manipulation zurückgegriffen wird, ist davon abhängig, was der Hauptgrund der Nutzung eines Video Deepfakes ist.[5]

Face Swapping

Eine der Arten ist das Face Swapping, bei dem die Gesichter auf Bildern oder Videos durch Fake Gesichter oder Gesichter anderer Personen, wie zum Beipsiel eines Promis, ersetzt wird. Dadurch ist es möglich die Person, dessen Gesicht verwendet wird, in einen anderen Kontext darstellen zu lassen, um beispielsweise in der Filmindustrie den Schauspieler mit einem Stunt Double austauschen zu können, um bestimmte Actionszenen realistischer wirken zu lassen.[5]

Face Morphing

Die zweite Art von Video Deepfakes ist das Face Morphing, welches ein Spezialeffekt ist, um ein Bild oder eine Form durch einen nahtlosen Übergang in ein anderes verändern zu können. Dieser Effekt wird oft in Filmen oder Animationen verwendet.[5]

Reenactment

Reenactment (auch face transfer or puppeteering genannt) ist eine Technik, bei der die Gesichtsausdrücke eines Source Images entsprechend eines Target Images angepasst werden. Das bedeutet, das die Gesichtsausdrücke und (Lippen-, Augen-, usw.) Bewegungen der Ursprungsperson mit denen einer anderen Person ersetzt werden.

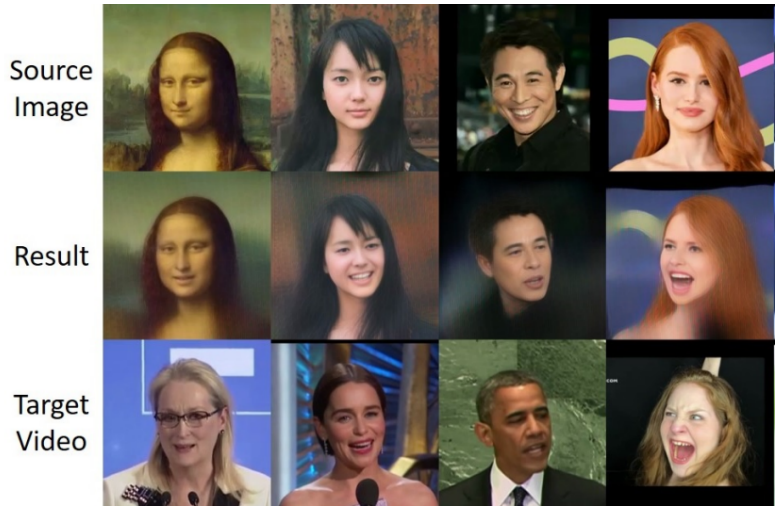

Abbildung 1.1: Reenactment Ergebnisse eines FSGAN[6]

Full body puppetry

Die letzte Art von Video Deepfakes ist die Full body puppetry, bei der einzelne Bewegungen bis hinzu komplette Bewegungsabläufe auf eine andere Person übertragen werden.
Die meisten Deepfakes benötigen viel Zeit für die Erstellung aufgrund der Systeme, welche erst mit dem Ausgangsmaterial trainiert werden müssen, um danach Inhalte verändern zu können. Es gibt aber auch Deepfake-Methoden die in Echtzeit funktionieren, welche die Möglichkeit bietet, Mimik und Lippenbewegungen einer Person zu erkennen und diese anschließend in Echtzeit auf das Videobild einer anderen Person übertragen zu lassen.[5]

Audio Deepfakes

Eine andere Art von Deepfakes sind Audio Deepfakes, bei dem aufgenommene oder live Audio Dateien verändert werden. Wobei hier zwischen Voice Swapping und Text to Speech unterschieden wird.[5]

Voice-swapping

Bei dem Voice-swapping können Audioinhalte so verändert werden, dass ein Text von einer fremden Person gesprochen werden kann. Die Stimme kann mit verschiedenen Effekten verändert werden, sodass zum Beispiel eine Stimme jünger, älter, männlich, weiblich oder auch mit verschiedenen Dialekten versehen werden kann. Dadurch wird dem Höhrer vorgespielt, dass verschiedene Personen sprechen, wobei es sich aber nur um eine Person hält.[5]

Text to Speech

Beim Text to Speech können Audioinhalte einer Aufnahme durch Eingabe eines neuen Textes verändert werden. Dadruch können zum Beispiel falsch ausgesprochene Wörter im nachhinein ersetzt werden, ohne eine neue Aufnahme durchführen zu müssen.[5]

Foto Deepfakes

Die dritte Art der Deepfakes sind Foto Deepfakes, bei denen es sich darum handelt, Fotos zu manipulieren. Dadurch können Fotos nach belieben verändert werden, um beispielsweise eine Person auf dem Bild durch einen Alterungsfilter, den Alterungsprozess der Person dargestellt werden kann.[5]

Face and body-swapping

Mithilfe des Deepfake-Algorithmus, welcher auch bei den anderen Arten verwendet wird, können Änderungen an einem Gesicht und Körper gemacht werden, indem das Gesicht oder der Körper mit einer anderen Person ausgetauscht wird. Eine mögliche Anwendung hierfür wäre das virtuelle anprobieren einer Brille, Haarfarbe oder Kleidung.[5]

Kombination aus Audio und Video Deepfake

Zuletzt gibt es wie oben eine mögliche Kombination der verschiedenen Arten, wie zum Beispiel die Kombination aus Audio und Video Deepfake. Diese Kombination wird auch das Lip-syncing genannt, bei dem Mundbewegungen sowie die gesprochenen Wörter in einem Video verändert und synchronisiert werden. Dadurch ist es möglich eine Person in einem Video scheinbar etwas sagen zu lassen, was sie aber niemals gesagt hat. Dies kann sowohl stark Missbraucht werden, indem zum Beispiel einem Politiker eine falschaussage untergeschoben wird. Es kann aber auch für positive Sachen Verwendung finden, um beispielsweise einen Film oder Werbung in eine andere Sprache zu synchronisieren.[5]

1.1.5 Anwendungsgebiete

Positive Anwendungsgebiete

Kunst- und Filmbranche

Negative Anwendungsgebiete

Politik und Regierung

Wirtschaft

Erstellung künstlicher Indentitäten

Mobbing

Pornographie

1.1.6 Ethik

1.2 Social Engineering

Unter Social Engineering werden alle Angriffe zusammengefasst, die die Schwachstelle Mensch ausnutzen. Es wird durch verschiedene Techniken versucht, an private oder sensible Inhalte von Personen zu gelangen. Social Engineering ist heutzutage eine der größten Gefahren im digitalen Raum. Kryptografische Verfahren und Protokolle wurden über die Jahre immer besser. Ist ein System bzw. eine digitale Infrastruktur richtig gehärtet, sind Angriffe wie Brute-Force oder Dictionary-Attacks wirkungslos. Außerdem werden durch moderne IDSs (Intrusion Detection Systems) sowie Threat Intelligence technische Angriffe immer schneller erkannt und blockiert. Laut einem Paper aus 2018 sind 84% aller Cyber-Angriffe auf Social Engineering zurückzuführen. Zudem haben Social Engineering Angriffe eine höhere Erfolgschance als herkömmliche Methoden. Laut des DBIR (Data Breach Investigations Report) von Verizon waren 2024 45% der erfassten Cyberangriffe auf Social Engineering zurückzuführen. Bei 83% der 3661 reporteten Incidents wurden Daten extrahiert.[7]

1.2.1 Verschiedene Typen von Social Engineering

Social Engineering lässt sich in verschiedene Bereiche untergliedern, die folgenden Beschreibungen richten sich nach dem Artikel: “Social Engineering Attacks: A Survey”[8]

Hier werden Social Engineering Angriffe in zwei Kategorien unterteilt.

- Human-based

Diese Angriffe werden manuell von einem Menschen ausgeführt. Sie sind in der Regel spezifisch auf das Opfer angepasst und mit höherem Aufwand verbunden. Dafür sind die Erfolgschancen höher als bei automatisierten Angriffen.

- Computer-based

Diese Angriffe werden automatisiert durchgeführt. Sie sind qualitativ deutlich schlechter als ihr Gegenstück, dafür werden sie in hoher Quantität durchgeführt. Hierzu zählen Phishing-Mails oder SMS. Es gibt verschiedene Tools, um solche Angriffe durchzuführen, ein bekanntes ist das “Social Engineering Toolkit”.

Des Weiteren können Social Engineering Angriffe in drei weitere Kategorien unterteilt werden.

- Social-based

Diese Form von Social Engineering Angriffen besteht aus zwischenmenschlicher Interaktion. Dabei spielt sie mit der Psychologie und den Emotionen der Zielperson. Diese Form von Social Engineering birgt ein hohes Risiko, hat aber ebenfalls eine hohe Erfolgschance, da der Angreifer im direkten oder indirekten Kontakt mit dem Opfer steht. Beispiele hierfür wären: Baiting, Spear-Phishing, aber auch Dinge wie der Enkeltrick.

- Technical-based

Hier werden Angriffe übers Internet remote ausgeführt. Dafür werden Social-Media Plattformen und Online-Dienste verwendet, um Passwörter, Kreditkarteninformationen oder personenbezogene Daten zu stehlen. Hierzu zählen zum Beispiel Phishing-Kampagnen oder gefälschte Webseiten.

- Physical-based

Physical-based Angriffe geschehen abseits des Internets in der realen Welt. Dabei werden durch physisches Handeln Informationen erschlossen. Ein Beispiel wäre das Durchsuchen von Müllcontainern (auch Dumpster-Diving genannt) nach sensiblen Dokumenten.

Je nachdem, aus welchem Blickwinkel die verschiedenen Techniken des Social Engineerings betrachtet werden, können diese in noch mehr verschiedene Kategorien eingeteilt werden. Neben Human-, Computer-, Social-, Technical- und Physical-based Social Engineering ist die zusätzliche Unterscheidung in **direkt** und **indirekt** sinnvoll. Ersteres benötigt direkten Kontakt zwischen Angreifer und Opfer, dabei zählen physischer Kontakt sowie

Telefonate. Beispiele sind: physical access, shoulder surfing, dumpster diving, phone social engineering, pretexting, impersonation on help desk call. Indirekte Angriffe sind entsprechend analog dazu. Hierzu zählen: phishing, fake software, Pop-Up windows, ransomware, SMSishing, online social engineering.

1.2.2 Überblick über gängige Angriffe

Phishing ist eine der am weitesten verbreiteten Social Engineering-Techniken. Ziel dieser Angriffe ist es, private oder vertrauliche Daten der Opfer zu stehlen. Dabei werden hauptsächlich E-Mails, SMS, Anrufe oder gefälschte Webseiten eingesetzt, um die Opfer zur Preisgabe ihrer Informationen zu verleiten. Phishing lässt sich grob in folgende Kategorien unterteilen[9]:

- **Spear Phishing**: Diese Methode ist zielspezifisch und verwendet oft durch Open Source Intelligence (OSINT) gesammelte Informationen, um maßgeschneiderte E-Mails zu erstellen. Die Nachrichten wirken dadurch besonders glaubwürdig und erhöhen die Erfolgschancen des Angriffs.
- **Whaling Phishing**: Hierbei handelt es sich um Angriffe auf hochrangige Ziele, wie Führungskräfte oder Personen in Schlüsselpositionen. Diese Angriffe sind oft sehr aufwendig und spezifisch auf das Ziel zugeschnitten, um wertvolle Informationen zu erlangen.
- **Vishing**: Voice Phishing, bei dem Telefonanrufe oder Sprachdienste wie Teams genutzt werden, um sensible Informationen zu erlangen. Die Angreifer geben sich häufig als vertrauenswürdige Institutionen oder Personen aus, um das Vertrauen des Opfers zu gewinnen.

Eine weitere Technik ist **Baiting**. Baiting, auch als Road Apples bekannt, verleitet Personen dazu, auf etwas zu klicken oder ein Gerät zu benutzen, um vermeintlich etwas gratis zu erhalten. Ein bekanntes Beispiel hierfür sind E-Mails mit einem Gewinn, für den man sich nur noch registrieren braucht, um ihn zu erhalten. Außerdem gehören auch infizierte USB-Sticks, die in der Hoffnung verteilt werden, dass jemand sie benutzt, zu Baiting dazu. Bei Bad-USBs wird auf die Neugierde des Menschen gesetzt. Durch das Einstecken des USB-Sticks in einen Computer kann Schadsoftware installiert werden, die es den Angreifern ermöglicht, auf das System zuzugreifen.

Tailgating Attacks beziehen sich auf das unerlaubte Verschaffen von Zutritt zu gesicherten Bereichen, indem zum Beispiel einer autorisierten Person gefolgt wird. Auch Angriffe auf die Sicherheitsmechanismen, wie z.B. das Kopieren eines NFC- oder RFID-Tags gehören in diese Kategorie. Solche Angriffe ermöglichen es dem Angreifer, physisch gesicherte Bereiche zu betreten und dort Informationen zu stehlen oder Schaden anzurichten.

1.2.3 Gegenmaßnahmen gegen Social Engineering Angriffe

Die Abwehr von Social Engineering Angriffen erfordert eine Kombination aus präventiven und reaktiven Maßnahmen. Eine der effektivsten Präventionsstrategien ist die **Schulung der Mitarbeiter**. Durch regelmäßige Schulungen und Sensibilisierungsprogramme können Mitarbeiter lernen, die Anzeichen von Social Engineering Angriffen zu erkennen und angemessen darauf zu reagieren. Dies umfasst das Überprüfen der Authentizität und Integrität von Nachrichten, sei es per E-Mail, SMS oder Telefon.

Überprüfung der Authentizität und Integrität von Nachrichten: Mitarbeiter sollten stets darauf achten, ungewöhnliche oder verdächtige Nachrichten sorgfältig zu prüfen. Dazu gehört, den Absender zu überprüfen, auf Rechtschreib- und Grammatikfehler zu achten und Links sowie Anhänge nicht ohne weiteres zu öffnen. Wenn Zweifel bestehen, sollte die Nachricht direkt beim vermeintlichen Absender verifiziert werden.

Da human-basierte Social Engineering Angriffe schwer oder gar nicht automatisiert zu erkennen sind, ist die **Schadensbegrenzung** von entscheidender Bedeutung. Hier kommen verschiedene technische und organisatorische Maßnahmen ins Spiel:

- **Domain-Tiering**: Diese Technik hilft, die Auswirkungen eines erfolgreichen Angriffs zu minimieren, indem unterschiedliche Sicherheitsstufen für verschiedene Domänen innerhalb eines Unternehmens festgelegt werden. Dadurch kann ein kompromittierter Bereich isoliert und der Schaden begrenzt werden.
- **Notfallmanagement**: Ein effektives Notfallmanagement umfasst klare Protokolle und Verantwortlichkeiten für den Fall eines Angriffs. Regelmäßige Schulungen und Übungen stellen sicher, dass alle Mitarbeiter wissen, wie sie im Ernstfall reagieren müssen. Dies beinhaltet auch die schnelle Identifikation und Isolation kompromittierter Systeme sowie die Benachrichtigung betroffener Personen und Behörden.

Zusätzlich zu diesen Maßnahmen können technische Hilfsmittel den Schutz vor Social Engineering Angriffen verbessern:

- **E-Mail-Sicherheitslösungen**: Tools wie E-Mail-Filter und Anti-Phishing-Software können verdächtige Nachrichten erkennen und blockieren, bevor sie die Mitarbeiter erreichen.
- **Zwei-Faktor-Authentifizierung (2FA)**: Durch die Implementierung von 2FA wird ein zusätzlicher Schutzlayer hinzugefügt, der es Angreifern erschwert, Zugang zu sensiblen Systemen und Daten zu erlangen, selbst wenn sie die Anmeldedaten eines Mitarbeiters gestohlen haben.
- **Netzwerküberwachung und IDSs**: Diese Systeme überwachen den Netzwerkverkehr auf verdächtige Aktivitäten und können Angriffe frühzeitig erkennen und abwehren.

Ein ganzheitlicher Ansatz, der sowohl präventive als auch reaktive Maßnahmen umfasst, ist unerlässlich, um die Widerstandsfähigkeit eines Unternehmens gegenüber Social Engineering Angriffen zu erhöhen. Durch die Kombination aus regelmäßiger Mitarbeiterschulung, technischer Absicherung und einem robusten Notfallmanagement kann das Risiko solcher Angriffe erheblich reduziert werden.[8], [10]

1.3 Bekannte Social Engineering Angriffe

Reenacted Video Call in Hong Kong

Ein besonders eindrucksvolles Beispiel für die Nutzung von Social Engineering in Verbindung mit Deepfake-Technologie ereignete sich Anfang des Jahres bei einem Unternehmen in Hongkong. Ein Finanzmitarbeiter wurde von Betrügern dazu gebracht, $25 Millionen zu überweisen. Wie die Hongkonger Polizei berichtet, wurde der Mitarbeiter zu einem Videoanruf eingeladen, jede Person in diesem Meeting war jedoch eine Deepfake. Mittels Echtzeit Face- und Voice-Reenactment wurde das gesamte Meeting gefälscht.

Der Betrug begann mit einer Nachricht, die angeblich vom Finanzchef des Unternehmens in Großbritannien stammte und von einer geheimen Transaktion sprach. Obwohl der Mitarbeiter zunächst misstrauisch war und einen Phishing-Versuch vermutete, ließ er sich durch die anschließende Videokonferenz überzeugen, da die anwesenden Personen wie seine tatsächlichen Kollegen aussahen und klangen.

Dieser Fall zeigt eindrucksvoll, wie fortschrittlich und gefährlich Deepfake-Technologie inzwischen ist. Sie wurde hier nicht nur verwendet, um ein realistisches Video der Zielperson zu erstellen, sondern auch, um mehrere scheinbar authentische Teilnehmer an einem Videoanruf zu simulieren. Die Täuschung flog erst auf, als der Mitarbeiter nachträglich beim Hauptsitz des Unternehmens nachfragte[11].

Marriott-Hotel Databreach

Ein weiteres Beispiel für einen Social Engineering Angriff ereignete sich im Juni 2022 im Marriott-Hotel am Flughafen von Baltimore im US-Bundesstaat Maryland. Kriminelle hatten sich mittels Social Engineering durch einen Mitarbeiter des Hotels Zugang zum Netzwerk verschafft und 20 GB an Daten abgeschöpft, darunter auch Kreditkartendaten von Gästen und interne Geschäftsdaten des Hotels. Marriott hat die Strafverfolgungsbehörden eingeschaltet und wird nach eigenen Abgaben etwa 400 Personen benachrichtigen. Eine Lösegeldforderung der Erpresser lehnte die Hotelkette ab[12].

Phishing-Kampanie United States Department of Labor

Ein weiteres Beispiel für einen Phishing-Angriff ist eine Kampagne aus dem Jahre 2022, die die United States Department of Labor (DoL) imitiert und Empfänger auffordert, Angebote einzureichen, um Office 365-Anmeldeinformationen zu stehlen. Die E-Mails passierten gekaperte Server, die gemeinnützigen Organisationen gehören, um E-Mail-Sicherheitsblöcke zu umgehen. Außerdem wurde die Sender Adresse gespooft um den tatsächlichen Domains des DoL zu entsprechen.

Die Angreifer geben sich als leitender DoL-Mitarbeiter aus, der den Empfänger einlädt, sein Angebot für ein laufendes Regierungsprojekt einzureichen. Die E-Mails enthalten einen gültigen Briefkopf, professionell gestalteten Inhalt und einen dreiseitigen PDF-Anhang mit einem scheinbar legitimen Formular.

Das PDF enthält eine „BID“-Schaltfläche, die die Opfer auf eine Phishing-Seite weiterleitet. Die gefälschte Seite sieht überzeugend aus und enthält identisches HTML und CSS wie die echte. Die Bedrohungsakteure haben sogar eine Anweisungs-Pop-up-Nachricht hinzugefügt, um die Opfer durch (Phishing-)Prozess zu führen.

Wer für ein Projekt bietet, wird zu einem Formular zur Erfassung von Anmeldeinformationen weitergeleitet, das die E-Mail-Adresse und das Passwort von Microsoft Office 365 der Opfer abfragt[13].

2. Deepfake Varianten

2.1 Video-Deepfakes

Video-Deepfakes sind eine besondere Form von gefälschten Medieninhalten, bei denen Gesichtsbilder in Videos so manipuliert werden, dass sie authentisch erscheinen. Diese Technologie hat erhebliche Auswirkungen auf die Bereiche Sicherheit, Verifikation und Social Engineering. Im Folgenden werden verschiedene Methoden zur Erstellung von Video-Deepfakes beschrieben und ihre technischen Details, Unterschiede, Gemeinsamkeiten sowie Vor- und Nachteile diskutiert.

2.1.1 Veraltete Methode: 3D Modellierung

Die frühe Methode zur Erstellung von Video-Deepfakes basierte auf der 3D-Modellierung. Hierbei wird ein dreidimensionales Modell des Gesichts einer Person erstellt, das anschließend animiert und in ein Video eingefügt wird.

Technische Details:

- Erstellung eines 3D-Gesichtsmodells durch Scannen oder Fotogrammetrie.
- Animierung des Modells basierend auf Zielausdrücken oder Bewegungsdaten.
- Einfügung des animierten Modells in das Zielvideo mithilfe von Tracking- und Rendering-Techniken.

Vor- und Nachteile:

- **Vorteile:** Hohe Kontrolle über die Gesichtsausdrücke und -bewegungen, gute Qualität bei statischen Szenen.
- **Nachteile:** Hoher Aufwand bei der Modellierung und Animation, Schwierigkeiten bei der realistischen Darstellung von dynamischen Szenen und feinen Details [14].

Einsatz im Kontext von Social Engineering: Aufgrund des hohen Aufwands und der technischen Expertise, die für die Erstellung benötigt wird, sind 3D-Modellierungs-Deepfakes weniger verbreitet in Social Engineering-Angriffen. Außerdem müssen die neuronalen Netze auf jedes Ziel- sowie Quellgesicht mit einigen Daten (Minuten Video) neu trainiert werden. Deshalb und aus Gründen der Performance ist dieser Ansatz für Echtzeitanwendungen, wie Videoanrufen, nicht geeignet [8].

2.1.2 Generative Adversarial Networks (GANs)

Mit dem Aufkommen von Generative Adversarial Networks (GANs) hat sich die Erstellung von Deepfakes erheblich vereinfacht und verbessert. GANs bestehen aus zwei neuronalen Netzwerken, einem Generator und einem Diskriminator, die gegeneinander trainiert werden.[15]

Technische Details:

- Der **Generator** erzeugt Bilder, die versuchen, echte Bilder nachzuahmen.
- Der **Diskriminator** bewertet die Bilder des Generators und unterscheidet zwischen echten und generierten Bildern.
- Durch diesen Wettbewerb verbessert sich die Qualität der erzeugten Bilder stetig.

Vor- und Nachteile:

- **Vorteile:** Hohe Qualität der erzeugten Bilder, Möglichkeit zur Erstellung realistischer und dynamischer Gesichtsausdrücke.
- **Nachteile:** Erfordert große Datenmengen und Rechenressourcen für das Training, anfällig für Artefakte und Unstimmigkeiten bei komplexen Bewegungen [14].

Einsatz im Kontext von Social Engineering: GAN-basierte Deepfakes sind effektiver und leichter zu erstellen, was sie zu einem mächtigen Werkzeug für Angreifer im Bereich Social Engineering macht. Sie können verwendet werden, um gefälschte Videos zu erstellen, die Vertrauen erwecken und die Opfer täuschen. GANs sind ebenfalls zu ineffizient in der Implementierung, um in Echtzeitanwendungen Gebrauch zu finden.

2.1.3 FSGAN und FSGANv2

Eine Weiterentwicklung der GAN-Technologie sind die Face Swapping GANs (FSGAN) und ihre verbesserte Version, FSGANv2. Diese Technologien sind speziell für den Gesichtstausch und die Gesichtsnachstellung entwickelt worden.

Technische Details:

- FSGAN nutzt GANs, um Gesichtszüge von einer Quelle auf ein Zielvideo zu übertragen, ohne dass eine explizite 3D-Modellierung erforderlich ist.
- FSGANv2 verbessert diese Methode durch bessere Algorithmen zur Anpassung der Gesichtsausdrücke und -bewegungen sowie durch die Verwendung von fortschrittlichen Netzwerktechniken, um die Konsistenz und Realismus zu erhöhen [16].

Vor- und Nachteile:

- **Vorteile:** Sehr realistische Ergebnisse, weniger Training und Daten erforderlich im Vergleich zu reinen GANs, bessere Anpassung an unterschiedliche Gesichtsausdrücke und Beleuchtungen.
- **Nachteile:** Trotz Verbesserungen immer noch anfällig für subtile Unstimmigkeiten, die bei genauer Betrachtung auffallen können [6].

Einsatz im Kontext von Social Engineering: FSGAN und FSGANv2 sind äußerst effektiv für Social Engineering-Angriffe, da sie hochrealistische Deepfakes erstellen können, die schwer zu erkennen sind. Sie können verwendet werden, um falsche Identitäten zu erstellen und Vertrauen zu gewinnen, was das Risiko und den potenziellen Schaden solcher Angriffe erhöht [17].

2.1.4 Gemeinsamkeiten und Unterschiede

Gemeinsamkeiten:

- Alle Methoden zielen darauf ab, realistische Fälschungen zu erstellen, die schwer zu erkennen sind.
- Nutzung von KI und maschinellem Lernen zur Verbesserung der Qualität und Realismus der erzeugten Videos.

Unterschiede:

- 3D-Modellierung erfordert manuelle Arbeit und ist weniger flexibel bei dynamischen Szenen.
- GANs und deren Weiterentwicklungen (FSGAN, FSGANv2) bieten automatisierte Lösungen mit höherer Qualität und Flexibilität.
- FSGAN und FSGANv2 sind speziell auf Gesichtsmanipulation optimiert und bieten bessere Ergebnisse bei der Anpassung an verschiedene Bedingungen[15].

Insgesamt zeigt sich, dass die Weiterentwicklung der Deepfake-Technologien, insbesondere durch GANs und deren Spezialisierungen wie FSGAN und FSGANv2, erheblich zur Verbesserung der Qualität und Realismus beigetragen hat. Dies stellt jedoch auch eine größere Bedrohung im Bereich des Social Engineering dar, da die Täuschungsabsicht hinter den erzeugten Medieninhalten immer schwerer zu durchschauen ist.

2.2 Face Swapping vs. Reenactment

Es gibt verschiedene Techniken von Video Deepfakes, im Folgenden werden Face-Swapping, sowie Reenactment näher betrachtet. Beide Varianten können mit den oben vorgestellten Möglichkeiten realisiert werden. Es gibt Modelle die für einen von beiden Anwendungsfällen besser geeignet sind, grundlegend basieren aber heutige Modelle immer auf GANs.

2.2.1 Face Swapping

Face Swapping, eine weit verbreitete Technik innerhalb der Deepfake-Technologie, beinhaltet das Austauschen eines Gesichts in einem Bild oder Video durch das Gesicht einer anderen Person. Diese Methode hat insbesondere in den letzten Jahren erhebliche Fortschritte gemacht, vor allem durch die Entwicklung von GANs. Bei Face Swapping wird das Gesicht der Zielperson durch ein anderes Gesicht ersetzt, wobei Merkmale wie Hautfarbe, Beleuchtung und Gesichtsausdrücke so angepasst werden, dass das Ergebnis möglichst realistisch wirkt. Diese Technik findet vor allem Anwendung in der Gestaltungs- und Medienbranche. Es können z.B. Gesichter von Schauspielern auf ihre Stuntdoubles gesetzt werden, um realistischere Stunt Szenen zu erzeugen. Eine besondere Form des Face Swapping ersetzt speziell den Mund eines Schauspielers, um die Synchronisation in anderen Sprachen zu vereinfachen[14]. In der Cyber-Security spielt diese Technik keine große Rolle, da nur das Gesicht ersetzt wird, müssen Dinge wie Hintergrund, Frisur und Kleidung selbst an die Zielperson angepasst werden. Dieser Aufwand ist heutzutage nicht mehr nötig, da Reenactmentmodelle ähnlich gute Ergebnisse erzielen.

2.2.2 Reenactment

Reenactment, auch als Face Transfer oder Puppeteering bekannt, ist eine Technik, bei der die Gesichtsausdrücke und -bewegungen eines Ausgangsbildes oder -videos auf ein Zielbild oder -video übertragen werden. Dies ermöglicht es, das Gesicht der Zielperson so zu manipulieren, dass es die gleichen Bewegungen und Ausdrücke wie das Ausgangsgesicht zeigt. Diese Technik findet sich ebenfalls in der Filmbranche wieder, indem z.B. verstorbene oder anderweitig verhinderte Schauspieler trotzdem noch in Filmen oder Serien zu sehen sind. Das vermutlich bekannteste Beispiel hierfür ist die Nutzung von Deepfake-Technologie, um die Charaktere Grand Moff Tarkin und Prinzessin Leia in “Rogue One: A Star Wars Story” realistischer darzustellen. In der originalen Filmproduktion wurden von beiden Charakteren alte 3D-Modelle bzw. Facescans verwendet um mit herkömmlichen Animations- und Rendertechniken realisiert. Durch den Einsatz von Face-Swapping und Reenactment wurden die visuellen Effekte dieser Charaktere von Fans so verbessert, dass sie natürlicher und lebensechter wirken als die ursprünglichen Effekte. Dieses Beispiel zeigt die Vorteile vom Einsatz von Deepfakes in der Filmproduktion und erweitern die Möglichkeiten der Branche erheblich[18].
Im Kontext von Cyber-Security sind die Anwendungsfälle offensichtlich. Es können durch Reenactment Videos von einflussreichen Personen innerhalb von Firmen erstellt werden, um Phishing noch effektiver zu gestalten. Außerdem können Videos von Personen des öffentlichen Lebens angefertigt werden in denen diese kontroverse Aussagen tätigen, um die öffentliche Meinung ins Negative zu ziehen. Ein gutes Beispiel hierfür ist “You Won’t Believe What Obama Say In This Video”.
Vor allem durch den Einsatz von auf Performance spezialisierter GANs können Videos nahezu in Echtzeit gefaked werden. Für Social Engineering relevante Videomedien sind ohnehin Video-Calls – dies hat zur Folge, dass ein Delay von wenigen Sekunden, sowie

kleine Artefakte oder Bildrauschen nicht ins Gewicht fallen, da diese auch von der Streamingplattform ausgehen könnten. Die Qualität der Deepfakes braucht also nicht auf filmreifen Niveau zu sein, um für Social Engineering brauchbar zu sein. Dies hat zur Folge, dass schon mit wenig Aufwand und Wissen, eine großzahl von Personen solche Fakes erstellen können.

2.3 Detection Methods für Video-Deepfakes

Die rasante Verbreitung von Deepfake-Inhalten stellt eine erhebliche Bedrohung für die Privatsphäre, die soziale Sicherheit und die Integrität des Internets dar. Um diesen Bedrohungen entgegenzuwirken, sind effektive Erkennungsmethoden unerlässlich. Verschiedene Techniken wurden entwickelt, um die Authentizität von Videos zu prüfen und Deepfakes zu identifizieren.

Temporale Sequenzanalyse

Eine der gängigsten Methoden zur Erkennung von Deepfakes ist die temporale Sequenzanalyse. Diese Technik nutzt die Fähigkeit von LSTM (Long Short-Term Memory) Netzwerken und CNNs (Convolutional Neural Networks), um zeitliche Unstimmigkeiten zwischen den Frames eines Videos zu erkennen. Durch die Analyse der Sequenzen von Frames können LSTM-Netzwerke zusammen mit CNNs Muster identifizieren, die auf Deepfake-Manipulationen hinweisen. Hierbei extrahieren die CNNs eine Vielzahl von Merkmalen aus jedem Frame und übergeben diese an die LSTM-Netzwerke, die eine temporale Sequenzbeschreibung erzeugen. Eine SoftMax-Schicht berechnet schließlich die Wahrscheinlichkeit, dass die analysierten Frames Deepfakes sind[14].

Blinzelmuster Erkennung

Eine weitere Methode basiert auf der Analyse der Augenblinzelmuster in Videos. Deepfake-Videos weisen oft unnatürliche Blinzelraten auf, da das Blinzeln in den synthetisierten Videos schlecht dargestellt wird. Hierfür wird das Video in einzelne Frames konvertiert und die Augenbereiche werden extrahiert. Diese Augenbereich-Sequenzen werden durch LRCNs (Long-Term Recurrent Convolutional Networks) verarbeitet, um die Wahrscheinlichkeit der Augenöffnungs oder -schließzustände vorherzusagen. Diese Methode ist besonders effektiv, da menschliche Blinzelmuster schwer nachzuahmen und synthetisierbar sind[14].

Physiologisch basierte Erkennungsmethoden

Physiologisch basierte Erkennungsmethoden nutzen Unterschiede zwischen computergenerierten Gesichtern und realen menschlichen Gesichtern. Ein Beispiel hierfür ist die Analyse von Blutflussmustern im Gesicht, die in Deepfake-Videos oft fehlen oder unnatürlich dargestellt werden. Solche physiologischen Signale können mit spezieller Software erfasst und analysiert werden, was eine hohe Genauigkeit bei der Erkennung von subtilen Unterschieden zwischen echten und gefälschten Gesichtern ermöglicht[19].

Digitale Wasserzeichen und Blockchains

Digitale Wasserzeichen und Blockchain-Technologien bieten ebenfalls effektive Möglichkeiten zur Authentifizierung von Videoinhalten. Digitale Wasserzeichen können in Videos eingebettet werden und gehen bei Manipulationen verloren. Die Blockchain-Technologie kann verwendet werden, um digitale Signaturen zu speichern und die Verbreitung von Videos zu verfolgen. Dies bietet eine hohe Sicherheit und Transparenz bei der Verifizierung der Authentizität von Videoinhalten, erfordert jedoch einen hohen Implementierungsaufwand[19].

Echtzeiterkennung durch Augenspiegelung

Aktive forensische Methoden sind besonders nützlich für die Echtzeit-Erkennung von Deepfakes, beispielsweise bei Videoanrufen. Diese Methoden nutzen spezifische Muster oder Reize, um Deepfakes in Echtzeit zu erkennen. Ein Beispiel ist die Anzeige eines unverwechselbaren Musters auf dem Bildschirm und die Analyse der kornealen Reflexion im Auge des Gesprächspartners. Solche biometriebasierten Ansätze sind effektiv für die Echtzeit-Erkennung und bieten eine robuste Lösung zur Verhinderung von Deepfake-Angriffen in Videokonferenzen[20].

Insgesamt bieten die verschiedenen Erkennungsmethoden für Video-Deepfakes eine breite Palette von Ansätzen zur Identifizierung und Authentifizierung von Inhalten. Die kontinuierliche Weiterentwicklung dieser Technologien ist entscheidend, um den immer leistungsfähigen Deepfake-Techniken entgegenzuwirken.

3. Erstellung von Deepfake Videos

Um Deepfakes erstellen zu können, gibt es einige verschiedene Tools. Die verbreitetsten Tools sind DeepFaceLab für Video-Deepfakes und DeepFaceLive für JIT (Just-in-Time)-Anwendungen. Im Folgenden werden beide Tools mit ihren verschiedenen Möglichkeiten sowie Best-Practices vorgestellt.

3.1 DeepFaceLab

DFL (DeepFaceLab) ist ein Open-Source Framework zur Erstellung von Face-Swapping Videos. DFL pipelined den Prozess der fotorealistischen Videomanipulation.

3.1.1 Motivation

Die Motivation hinter DFL ist es, sowohl die Erstellung als auch die Erkennung von Deepfakes zu verbessern. Durch die Bereitstellung eines leistungsfähigen und flexiblen Werkzeugs für Gesichtsmanipulationen trägt es zur Weiterentwicklung der Forschung im Bereich der Medienfälschungserkennung bei. Es hilft dabei, qualitativ hochwertige Fälschungsdaten zu erzeugen, die für die Entwicklung robuster Erkennungsmodelle unerlässlich sind[17], [21].

3.1.2 Fähigkeiten

DeepFaceLab zeichnet sich durch mehrere Hauptmerkmale aus:

- **Bequemlichkeit**: Der gesamte Workflow von DeepFaceLab, einschließlich Datenverarbeitung, Modelltraining und Nachbearbeitung, ist darauf ausgelegt, so benutzerfreundlich und effizient wie möglich zu sein. Es bietet ein vollständiges CLI (Command Line Interface), die eine flexible Implementierung ermöglicht.
- **Breite technische Unterstützung**: Das Tool unterstützt Multi-GPU-Konfigurationen und die Verwendung von mehreren Threads zur Beschleunigung grafischer Operationen und Datenverarbeitung. Laut Paper können selbst auf einem Rechner mit nur 2GB VRAM erfolgreiche Gesichtsmanipulationen durchgeführt werden[21].
- **Erweiterbarkeit**: Die Architektur von DeepFaceLab ist modular aufgebaut, sodass einzelne Komponenten einfach austauschen werden können[21].
- **Skalierbarkeit**: DFL unterstützt durch verschiedene Tools die Verarbeitung von großen Datenmengen.[21].

3.1.3 Workflow

Der Workflow von DeepFaceLab ist in der Form einer Pipeline und besteht aus drei Hauptphasen: Extraction, Training und Conversion.

Extraction

In der Extraktionsphase werden Gesichter aus den Quell- und Zielvideos extrahiert. Diese Phase umfasst mehrere Algorithmen und Verarbeitungsschritte, dazu gehören face detection, face alignment und face segmentation. DFL bietet verschiedene Extraktionsmodi (z.B. half-face, full-face, whole-face), um den unterschiedlichen Anforderungen gerecht zu werden[17], [21].

- **Face Detection**: Hierbei wird ein CNN verwendet, um Gesichter in den Videoframes zu erkennen. Diese Detektion ist entscheidend, um die Position und Größe des Gesichts für die weiteren Verarbeitungsschritte zu bestimmen.
- **Face Alignment**: Nachdem die Gesichter erkannt wurden, werden sie durch Algorithmen zur Gesichtsangleichung normalisiert. Dies bedeutet, dass die Gesichter in eine einheitliche Position und Größe gebracht werden, was die Genauigkeit der späteren Schritte erhöht.
- **Face Segmentation**: In diesem Schritt werden die Gesichter von den restlichen Bildinformationen getrennt. Dies ermöglicht eine gezielte Bearbeitung der Gesichter ohne Beeinträchtigung des restlichen Bildes.

Training

Das Training ist die entscheidende Phase, in der ein Modell trainiert wird, um realistische Gesichtsmanipulationen zu erzeugen. DeepFaceLab verwendet zwei Hauptstrukturen: die DF-Struktur und die LIAE-Struktur. Dabei wird eine Mischung aus DSSIM- und MSE-Verlusten verwendet, um sowohl eine schnelle Generalisierung als auch eine hohe Präzesion zu erreichen[17], [21].

- **DF-Struktur (DeepFakes)**: Diese Struktur basiert auf GANs und nutzt zwei Netzwerke - ein Generator und ein Diskriminator. Der Generator versucht, realistische Gesichter zu erzeugen, während der Diskriminator versucht, zwischen echten und generierten Gesichtern zu unterscheiden. Durch diesen Wettbewerb lernen beide Netzwerke, immer realistischere Ergebnisse zu produzieren.
- **LIAE-Struktur (Learned Image-to-Image Translation)**: Diese Struktur verwendet ebenfalls GANs, fokussiert sich aber stärker auf die Übersetzung von Bildern. Es wird ein Encoder-Decoder-Ansatz verwendet, bei dem das Gesicht in einen latenten Raum kodiert und anschließend in das Zielgesicht dekodiert wird.
- **Verlustfunktionen**: Der DSSIM (Structural Dissimilarity) Verlust wird verwendet, um strukturelle Unterschiede zwischen dem generierten und dem echten Bild zu minimieren, während der MSE (Mean Squared Error) Verlust die pixelweisen Unterschiede minimiert.

Konvertierung

In der Konvertierungsphase werden die erzeugten Gesichter wieder in die ursprünglichen Zielbilder eingefügt. Dieser Schritt umfasst Farbanpassungen, um den Hautton und die Beleuchtung anzugleichen, sowie das Schärfen der Bilder, um Details hervorzuheben. DeepFaceLab bietet mehrere Farbanpassungsalgorithmen und nutzt ein vortrainiertes Super-Resolution-Netzwerk, um die endgültigen Bilder zu verbessern[17], [21].

- **Farbanpassung**: Hierbei werden Algorithmen verwendet, die den Hautton und die Beleuchtung des generierten Gesichts an das Zielbild anpassen, um einen nahtlosen Übergang zu gewährleisten.
- **Super-Resolution**: Ein vortrainiertes Super-Resolution-Netzwerk wird eingesetzt, um die Auflösung der generierten Gesichter zu erhöhen und feinere Details hervorzuheben. Dies ist besonders wichtig für hochauflösende Videos.
- **Nachbearbeitung**: Weitere Nachbearbeitungsschritte umfassen das Schärfen der Bilder und das Entfernen von Artefakten, um die Gesamtqualität zu verbessern.

3.2 Praxisbeispiel

Im Folgenden wird der Workflow mit DFL näher betrachtet. Anschließend werden verschiedenen Konfigurationen und Trainingsdauern verglichen.

3.2.1 Laborumgebung

Alle Deepfakes, wenn nicht anders genannt, werden auf folgender Hardware erstellt.

CPU:	AMD Ryzen 5 2600X
RAM:	16GB DDR4 3000MHz
GPU:	NVIDIA RTX 2070 (8GB GDDR6 VRAM)
OS:	Windows 11

Ziel des Deepfakes ist das Gesicht von RDJ (Robert Downey Jr.) auf Prof. Volker Knoblauch in diesem https://www.youtube.com/watch?v=rksMPlRSbQU zu swappen, sodass die Hochschule Aalen amerikanische Prominente auf ihrem Youtube-Kanal zeigen kann.

3.2.2 Extraction

Nach der Installation von DFL finden sich im entsprechenden Ordner eine Vielzahl von `.bat`-Dateien, sowie ein `_internal`- und `workspace`-Ordner.

Literaturverzeichnis

[1] CrowdStrike. (4. Juli 2024). „What is Cyber Threat Intelligence? [Beginner's Guide]“ [Online]. Verfügbar: https://www.crowdstrike.com/cybersecurity-101/threat-intelligence/

[2] M. Block. (29. Aug. 2023). „Definition und Anwendungsbereiche“ [Online]. Verfügbar: https://link.springer.com/chapter/10.1007/978-3-662-67427-7_2

[3] L. Whittaker. (Juli 2023). „Mapping the deepfake landscape for innovation: A multidisciplinary systematic review and future research agenda“ [Online]. Verfügbar: https://www.sciencedirect.com/science/article/pii/S0166497223000950#abs0015

[4] J. A. Marwan Albahar. (30. Nov. 2019). „DEEPFAKES: THREATS AND COUNTERMEASURES SYSTEMATIC REVIEW“ [Online]. Verfügbar: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.jatit.org/volumes/Vol97No22/7Vol97No22.pdf

[5] J.-T. Kötke. (Feb. 2021). „DEEPFAKE -EINE KURZE EINLEITUNG Deepfake -Eine kurze Einleitung“ [Online]. Verfügbar: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.researchgate.net/profile/Jennifer-Tia-Koetke/publication/373041489_DEEPFAKE_-EINE_KURZE_EINLEITUNG_Deepfake_-Eine_kurze_Einleitung/links/64d4ffddd3e680065aac7ee3/DEEPFAKE-EINE-KURZE-EINLEITUNG-Deepfake-Eine-kurze-Einleitung.pdf

[6] Y. Nirkin, Y. Keller und T. Hassner, „FSGAN: Subject Agnostic Face Swapping and Reenactment“, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Okt. 2019.

[7] C. D. Hylender, P. Langlois, A. Pinto und S. Widup, „2024 Data Breach Investigations Report“, Verizon, Apr. 2024, Available at https://www.verizon.com/business/resources/reports/dbir/.

[8] F. Salahdine und N. Kaabouch, „Social Engineering Attacks: A Survey“, *Future Internet*, Jg. 11, Nr. 4, 2019, ISSN: 1999-5903. DOI: 10.3390/fi11040089. [Online]. Verfügbar: https://www.mdpi.com/1999-5903/11/4/89.

[9] K. Krombholz, H. Hobel, M. Huber und E. Weippl, „Advanced social engineering attacks“, *Journal of Information Security and Applications*, Jg. 22, S. 113–122, 2015, Special Issue on Security of Information and Networks, ISSN: 2214-2126. DOI: https://doi.org/10.1016/j.jisa.2014.09.005. [Online]. Verfügbar: https://www.sciencedirect.com/science/article/pii/S2214212614001343.

[10] BSI. (5. Juli 2024). „Social Engineering - der Mensch als Schwachstelle“ [Online]. Verfügbar: https://www.bsi.bund.de/DE/Themen/Verbraucherinnen-und-Verbraucher/Cyber-Sicherheitslage/Methoden-der-Cyber-Kriminalitaet/Social-Engineering/social-engineering_node.html

[11] CNN. (5. Juli 2024). „Finance worker pays out $25 million after video call with deepfake ‘chief financial officer’“ [Online]. Verfügbar: https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html

[12] DSGVO-Portal. (5. Juli 2024). „Datenpanne bei Marriott International, Inc. | Sicherheitsvorfalls-Datenbank“ [Online]. Verfügbar: https://www.dsgvo-portal.de/sicherheitsvorfaelle/datenpanne_bei_marriott-international-inc.-1069.php

[13] B. Computer. (5. Juli 2024). „Office 365 phishing attack impersonates the US Department of Labor“ [Online]. Verfügbar: https://www.bleepingcomputer.com/news/security/office-365-phishing-attack-impersonates-the-us-department-of-labor/

[14] A. Chadha, V. Kumar, S. Kashyap und M. Gupta, *Deepfake: An Overview*, P. K. Singh, S. T. Wierzchoń, S. Tanwar, M. Ganzha und J. J. P. C. Rodrigues, Hrsg. Singapore: Springer Singapore, 2021, S. 557–566, ISBN: 978-981-16-0733-2.

[15] D. Cavedon-Taylor, „Deepfakes: a survey and introduction to the topical collection“, *Synthese*, Jg. 204, Nr. 1, S. 14, 2024, ISSN: 1573-0964. DOI: 10.1007/s11229-024-04634-8. [Online]. Verfügbar: https://doi.org/10.1007/s11229-024-04634-8.

[16] Y. Nirkin, Y. Keller und T. Hassner, „FSGANv2: Improved Subject Agnostic Face Swapping and Reenactment“, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Jg. 45, Nr. 1, S. 560–575, Jan. 2023, PubMed-not-MEDLINE, PMID: 35471874, ISSN: 1939-3539, 0098-5589. DOI: 10.1109/TPAMI.2022.3155571. [Online]. Verfügbar: https://doi.org/10.1109/TPAMI.2022.3155571.

[17] K. Liu, I. Perov, D. Gao, N. Chervoniy, W. Zhou und W. Zhang, „Deepfacelab: Integrated, flexible and extensible face-swapping framework“, *Pattern Recognition*, Jg. 141, S. 109 628, 2023, ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2023.109628. [Online]. Verfügbar: https://www.sciencedirect.com/science/article/pii/S0031320323003291.

[18] C. Blend. (7. Juli 2024). „Rogue One Deepfake Makes Star Wars’ Leia And Grand Moff Tarkin Look Even More Lifelike“ [Online]. Verfügbar: https://www.cinemablend.com/news/2559935/rogue-one-deepfake-makes-star-wars-leia-and-grand-moff-tarkin-look-even-more-lifelike

[19] M. Westerlund, „The Emergence of Deepfake Technology: A Review“, *Technology Innovation Management Review*, Jg. 9, S. 40–53, Nov. 2019, ISSN: 1927-0321. DOI: http://doi.org/10.22215/timreview/1282. [Online]. Verfügbar: timreview.ca/article/1282.

[20] H. Guo, X. Wang und S. Lyu, „Detection of Real-Time Deepfakes in Video Conferencing with Active Probing and Corneal Reflection“, in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, S. 1–5. DOI: 10.1109/ICASSP49357.2023.10094720.

[21] I. Perov, D. Gao, N. Chervoniy u. a., *DeepFaceLab: Integrated, flexible and extensible face-swapping framework*, 2021. arXiv: 2005.05535 [cs.CV]. [Online]. Verfügbar: https://arxiv.org/abs/2005.05535.