

# BACHELORARBEIT

## Der Forschungsstand zu Deepfakes und deren Erstellung

Eingereicht am 08. Juni 2021, Zürich

**Betreut von:**

Prof. Dr. Thomas Keller

**Student:**

Yves Wegmann



## Management Summary

Fortschritte im Bereich der künstlichen Intelligenz und neuronalen Netzwerken haben zur Generierung von realistischen gefälschten Inhalten geführt. Diese neue Technologie mit dem Namen Deepfake hat sich in den letzten Jahren enorm weiterentwickelt und eine grosse Vielfalt von gut- und bössartigen Anwendungen gefördert. Einfach zugängliche Tools bieten die Möglichkeit, jemanden darzustellen, der bestimmte Dinge sagt und tut, die nie passiert sind. In Kombination mit der Reichweite und Geschwindigkeit von sozialen Medien können Deepfakes in kürzester Zeit eine hohe Anzahl von Leuten erreichen. Rasant haben Wissenschaftler und Deepfake-Communities Forschungen zur Deepfake-Erzeugung und Deepfake-Erkennung durchgeführt, wobei jeweils die Verbesserungen der einen Seite die andere antreibt und neue Methoden entstehen. Dennoch ist der Überblick über diese neuen Ansätze sowie die Verbreitung von Deepfakes vernachlässigt worden.

Die vorliegende Arbeit untersucht den Forschungsstand zur Verbreitung und Erstellung von Deepfakes. Dabei werden anhand einer Literaturrecherche die verschiedenen Kategorien zur Erzeugung von Deepfakes vorgestellt und der Nutzen sowie die Bedrohungen, die sich aus den synthetischen Inhalten ergeben, diskutiert. Ebenso werden aktuelle Lösungsansätze zur Entdeckung und Bekämpfung von böswilligen Deepfakes vorgestellt. Mithilfe von zwei durchgeführten Studien kann die aktuelle Verbreitungslandschaft der neuen Technologie erläutert werden. Eine Artikelsuche von elf ausgewählten Zeitungen aus verschiedenen Regionen der Schweiz analysiert die grobe Berichterstattung zum Thema und bekräftigt die Annahme über ein geringes Verständnis von Deepfakes in der Schweizer Bevölkerung. Mit eigenem Videomaterial und einer der am häufigsten verwendeten Anwendung zur Erzeugung von Deepfakes wird ein eigenes synthetisches Video erstellt.

Insgesamt sind fünf verschiedene Erzeugungskategorien identifiziert sowie deren Einsatzmöglichkeiten und bestehende Herausforderungen erläutert worden. Die Studien zur Verbreitung von Deepfakes zeigen alle sechs Monate eine Verdoppelung der synthetischen Inhalte und liefern relevante Erkenntnisse über die Herkunft und Berufe der Opfer sowie den Einsatz der Deepfakes. Anhand der Erstellung eines eigenen gefälschten Inhaltes kann das Verständnis über den Aufwand und die einzelnen Arbeitsschritte verstärkt werden. Durch die aufgetretenen Herausforderungen sind zudem neue Ansätze zur Minderung dieser Probleme präsentiert worden. Die Literaturrecherche hat ausserdem einen

bedeutenden Einblick in die Nutzung und Bedrohungen sowie möglichen Lösungsmethoden zur Entdeckung und Bekämpfung von Deepfakes ergeben.

Die Arbeit zeigt auf, dass Deepfakes ohne Expertenkenntnisse erstellt werden können und ein grosses Forschungspotenzial im Bereich der Verbreitungslandschaft besteht. Der umfassende Einblick in die Generierung eines Deepfakes hilft ein Verständnis über die neue Technologie zu schaffen. Die Erkenntnisse im Bereich der Verbreitung sowie die fünf identifizierten Erzeugungskategorien können für weitere Forschungsarbeiten auf diesen Gebieten verwendet werden.

# Inhaltsverzeichnis

Inhaltsverzeichnis .....	IV
Abbildungsverzeichnis .....	VI
Tabellenverzeichnis .....	VI
Abkürzungsverzeichnis .....	VII
1 Einleitung.....	1
1.1 Ausgangslage.....	1
1.2 Forschungsfrage .....	1
1.3 Relevanz des Themas .....	2
1.4 Methodisches Vorgehen .....	3
2 Was ist ein Deepfake?.....	4
3 Technologische Grundlagen .....	5
3.1 Der Autoencoder .....	7
3.2 Generative Adversarial Networks (GANs) .....	9
3.3 Die verschiedenen Erzeugungsmethoden von Deepfakes.....	12
3.3.1 Face-Swap .....	12
3.3.2 Lip-Syncing.....	15
3.3.3 Puppet Master .....	17
3.3.4 Audio-Deepfakes.....	18
3.3.5 Gesichtssynthese und Merkmalmanipulation .....	20
3.4 Herausforderungen .....	22
4 Verbreitung von Deepfakes .....	23
4.1 Weltweite Landschaft von Deepfakes .....	24
4.2 Medialer Umfang über Deepfakes in der Schweiz.....	28
5 Nutzen und Bedrohungen von Deepfakes.....	34
5.1 Nutzen durch Deepfakes .....	35
5.1.1 Bildung.....	35
5.1.2 Kunst .....	36
5.1.3 Multimediaindustrie .....	37
5.2 Bedrohungen durch Deepfakes .....	38
5.2.1 KI-Pornos .....	39
5.2.2 Verfälschung der Realität.....	43
5.2.2.1 Politische Manipulationen.....	43
5.2.2.2 Gefahr für Unternehmen .....	45
5.2.2.3 Täuschung und Diffamierung von Einzelpersonen.....	47
6 Lösungen zur Bekämpfung von feindlichen Deepfakes .....	48
6.1 Technische Lösungen .....	49
6.2 Rechtliche Lösungen .....	50
7 Erstellung eines eigenen Deepfakes.....	53

7.1	Vorbereitung.....	53
7.1.1	Voraussetzungen .....	54
7.1.2	Auswahl der Open-Source-Software.....	54
7.1.3	Installation der Software .....	54
7.1.4	Videomaterial.....	55
7.1.5	Software kennenlernen.....	55
7.2	Erstellung des eigenen Deepfakes .....	58
7.2.1	Extrahierung der Quelle .....	58
7.2.2	Extrahierung des Zielvideos.....	59
7.2.3	Extrahierung der Gesichter «data_src» .....	59
7.2.4	Extrahierung der Gesichter «data_dst» .....	60
7.2.5	Training des Modells.....	61
7.2.6	Zusammenführen.....	62
7.2.7	Video erstellen .....	64
7.2.8	Erkenntnisse .....	65
7.3	Bearbeitung Okklusion und Posenschwankung .....	66
7.3.1	Mehr Iterationen.....	66
7.3.2	Neues Videomaterial.....	67
7.3.3	Training mit SAEHD .....	69
7.4	Zusammenfassung eigener Deepfake .....	71
8	Schlussteil .....	72
8.1	Beantwortung der Forschungsfrage.....	72
8.2	Beurteilung der Hypothesen .....	74
8.2.1	Hypothese 1.....	74
8.2.2	Hypothese 2.....	75
8.2.3	Hypothese 3.....	75
8.2.4	Hypothese 4.....	76
8.3	Fazit.....	77
	Literaturverzeichnis .....	78
	Anhang .....	89
A	Deepfake-Anleitung der Community .....	89

## **Abbildungsverzeichnis**

Abbildung 1: Architektur eines neuronalen Netzes (Zucconi, 2018).....	7
Abbildung 2: Ablauf eines Autoencoder (Deshmukh & Wankhade, 2021, S. 295).....	9
Abbildung 3: Die ersten GAN-Gesichtsbilder (Goodfellow et. al., 2014, S. 6).....	11
Abbildung 4: Heutige synthetische GAN-Bilder (This Person Does Not Exist, 2021).	11
Abbildung 5: Veranschaulichung des Reface-Prozesses.....	14
Abbildung 6: Erstellung Lippensynchronisation (Suwajanakorn et al., 2017, S. 2). ....	16
Abbildung 7: Kategorien von Gesichtsm Manipulationen (Masood et al., 2021, S. 12)....	21
Abbildung 8: Online gestellte Deepfakes (Cavalli, 2021).....	25
Abbildung 9: Entdeckter Deepfakes (Tammekänd et al., 2020, S. 7). ....	25
Abbildung 10: Akademische Arbeiten über GANs (Ajder et al., 2019, S. 9). ....	26
Abbildung 11: Unterschied der Geschlechter (Ajder et al., 2020, S. 8).....	27
Abbildung 12: Ein globales Phänomen (Ajder et al., 2020, S. 8). ....	27
Abbildung 13: Protagonisten und deren Berufe (Ajder et al., 2020, S. 8). ....	28
Abbildung 14: Anfrage in einem Deepfake-Forum.....	40
Abbildung 15: Generierung gefälschter Nacktbilder (Ajder et al., 2020, S. 4).....	42
Abbildung 16: Ordner in DFL .....	57
Abbildung 17: Erster Deepfake .....	57
Abbildung 18: Extrahierte Gesichter .....	61
Abbildung 19: Vorschaufenster beim trainieren .....	62
Abbildung 20: Bedienelemente in DFL.....	63
Abbildung 21: Eigener Deepfake .....	64
Abbildung 22: Posenschwankung und Okklusion.....	66
Abbildung 23: Ergebnisse mit neuem Videomaterial.....	69
Abbildung 24: Ergebnisse mit SAEHD.....	71

## **Tabellenverzeichnis**

Tabelle 1: Zeitungsartikel der deutschsprachigen Zeitungen. ....	30
Tabelle 2: Artikel der französischsprachigen Schweiz. ....	31
Tabelle 3: Artikel der italienischsprachigen Zeitungen.....	31
Tabelle 4: Artikel der Online-Nachrichtenportale.....	32
Tabelle 5: Alle Artikel nach Jahr.....	33

## Abkürzungsverzeichnis

AI	Artificial Intelligence
ALS	Amyotrophe Lateralsklerose
cGAN	Conditional Generative Adversarial Network
CGI	Computer Generated Imagery
CPU	Central Processing Unit
DCGAN	Deep Convolutional Generative Adversarial Network
DFL	DeepFaceLab
GANs	Generative Adversarial Networks
GIF	Graphics Interchange Format
GPU	Graphics Processing Unit
IcGAN	Invertible Conditional Generative Adversarial Network
KI	Künstliche Intelligenz
MOOC	Massive Open Online Course
RGB	Red Blue Green
WGAN	Wasserstein Generative Adversarial Network

# 1 Einleitung

Die vorliegende Bachelorarbeit beschäftigt sich mit dem aktuellen Forschungsstand zur Verbreitung und Erstellung von Deepfakes. In diesem Kapitel wird die Ausgangslage dargelegt sowie die Forschungsfrage, die Relevanz des Themas und das Vorgehen erläutert.

## 1.1 Ausgangslage

Manipulationen von Medien, Fotos, Video und Audio sind so alt wie die Medien selbst (Kietzmann et al., 2020, S. 135). Für die Veränderung von visuellen und auditiven Medien brauchte es bisher die Kenntnisse eines Spezialisten oder den Zugang zu immensen Ressourcen wie ein Hollywood-Studio oder eine nationale Regierung besitzt (Schick, 2020, S. 10). Der jüngste Auftritt von Deepfakes stellt jedoch einen Wendepunkt in der Erstellung von gefälschten Inhalten dar. Die technologischen Fortschritte im Bereich der künstlichen Intelligenz (KI) und des maschinellen Lernens ermöglichen jedem der einen Computer besitzt, gefälschte Videos zu erstellen, die von authentischen Medien sehr schwer zu unterscheiden sind (Westerlund, 2019, S. 39). Bereits in einigen Jahren wird dies sogar auf dem Smartphone möglich sein (Schick, 2020, S. 10). Somit hätte jeder die Macht, Menschen an Orten zu zeigen an denen sie nie gewesen sind, Dinge zu tun, die sie nie getan haben und Aussagen zu treffen, die sie nie gesagt haben. Angesichts der Leichtigkeit, mit der solche desinformativen Inhalte auf den Social-Media-Kanälen verbreitet werden können, wird es in Zukunft immer schwieriger werden zu wissen, welche Informationen korrekt sind (Mahmud & Sharmin, 2020, S. 19). Im kaputten Informations-Ökosystem, welches von Fehl- und Desinformationen geprägt ist, stellen KI und Deepfakes somit die neuste globale Bedrohung dar (Schick, 2020, S. 10).

## 1.2 Forschungsfrage

Das Hauptziel dieser Arbeit besteht darin, nach der dreijährigen Existenz von Deepfakes Bilanz zu ziehen und den aktuellen Stand zum Thema aufzuzeigen. Deshalb beschäftigt sich diese Arbeit mit folgender Forschungsfrage:

Wie ist der Forschungsstand zur Verbreitung und Erstellung von Deepfakes?



Folgende Hypothesen unterstützen dabei die Fragestellung:

- H1: Seit der Entdeckung von Deepfakes sind verschiedene Generierungsmethoden entwickelt worden.
- H2: In der Schweizer Gesellschaft herrscht ein zu geringes Verständnis über die neue Technologie.
- H3: Das Verhältnis zwischen den Nutzen und Bedrohungen durch Deepfakes ist nicht im Gleichgewicht.
- H4: Für die Bekämpfung von gefährlichen Deepfakes existieren einsatzfähige Lösungsansätze.

Die Forschungsfrage und die Hypothesen werden zusammen mit den Erkenntnissen aus der Arbeit im letzten Kapitel beantwortet und analysiert.

### **1.3 Relevanz des Themas**

Im Informationszeitalter verbreiten sich Videos und Bilder auf den Social-Media-Kanälen mit rasender Geschwindigkeit und erreichen in kürzester Zeit Millionen von Nutzern (Schick, 2020, S. 12). Bereits heute bezieht jeder fünfte Internetnutzer seine Nachrichten über YouTube und für das Jahr 2022 wird geschätzt, dass 82 Prozent des weltweiten Internetverkehrs aus Video-Streaming und Downloads stammen wird (Schick, 2020, S. 12). Folglich sind in Verbindung mit den jüngsten technologischen Entwicklungen im Bereich der KI die Möglichkeiten zur Täuschung dieser Inhalte endlos. Die Hürden und Anforderungen für die Erstellung von Deepfakes werden immer kleiner womit Milliarden von Nutzern nicht mehr nur Konsumenten sind, sondern auch selbst zu Produzenten werden (Schick, 2020, S. 37).

Dadurch dass heutzutage jeder Deepfakes erstellen kann, gibt es positive wie auch negative Optionen wie die neue Technologie genutzt wird (Pantserev, 2020, S. 41). Jedoch stellen Deepfakes in unserem Informationsökosystem, welches durch Fehl- und Desinformationen gekennzeichnet ist, die neuste sich entwickelnde Bedrohung dar (Schick, 2020, S. 12). Die bösertige Nutzung von Deepfakes kann eine Gefahr für Individuen, Konzerne sowohl aber auch für ganze Gesellschaften und Nationen darstellen.

Einige Experten sagten voraus, dass Entwickler bereits gegen Ende 2020 einen perfekten Deepfake erstellen können (Davis, 2020, S. 3). Ein hochentwickelter Deepfake ist frei von jeglichen Fehlern und wird von keinem Experten oder Algorithmus von echtem Filmmaterial unterschieden werden können (Davis, 2020, S. 3). Da ein grosser Teil des täglichen Wissens durch Videos angeeignet wird, besteht die grösste Bedrohung von Deepfakes darin, dass Menschen leicht dazu verleitet werden können falsche Überzeugungen zu glauben (Schick, 2020, S. 15). Deepfakes verändern somit die Art und Weise wie die Menschen über Lüge und Wahrheit denken. Es ist wichtig das Bewusstsein für die Existenz von Deepfakes in der Öffentlichkeit zu schärfen. Ebenso sollten die Wissenschaft, der Staat und vertrauenswürdige Institutionen neue Wege finden wie man Deepfakes schneller entdecken kann und Standards entwickeln, welche für wichtige Anwendungsfälle eingesetzt werden können.

#### **1.4 Methodisches Vorgehen**

In diesem Kapitel wird das methodische Vorgehen für die Beantwortung der Forschungsfrage beschrieben. Die Erarbeitung der Ergebnisse wird in drei Phasen eingeteilt. Als erstes wird ein Literaturreview durchgeführt. Dabei werden ausgewählte Datenbanken von Wissenschaftlichen Arbeiten sowie aktuelle Reports, Artikeln und Videos analysiert. Das Auswählen der relevanten Literatur wird durch die Suche mit definierten Stichwörtern sowie der Lesung des Abstracts erfolgen. Ausserdem werden explizite Deepfake-Foren und deren Aktivitäten analysiert.

In einem zweiten Schritt werden ausgewählte Zeitungen aus drei von vier Sprachregionen der Schweiz nach dem Stichwort «Deepfake» durchsucht. Die Suche wird über die eigene Suchoption der Zeitungs-Webseiten durchgeführt. Alle Artikel werden gesammelt, um eine empirische Analyse durchzuführen, wie die ausgewählten Zeitungen über Deepfakes berichtet haben. Dabei werden diese Publikationen mit dem Titel des Berichts sowie des Veröffentlichungsdatums erfasst. Anschliessend entsteht eine Auflistung der veröffentlichten Artikel, welche in vier Bereiche eingeteilt werden. Die Auswahl der Zeitungen orientiert sich nach dem WEMF-Auflagebulletin (WEMF AG, 2020). Durch diese Analyse kann eine Annahme über das Bewusstsein von Deepfakes in der Schweizer Bevölkerung getroffen werden. Die Ergebnisse der Wortsuche werden kritisch beurteilt.

Beim letzten Teil wird mittels einer bekannten öffentlichen Software (Open-Source-Software) ein eigener Deepfake erstellt. Die Arbeitsschritte für die Erzielung des optimalen Ergebnisses werden dokumentiert und in die Arbeit integriert. Das Vorgehen zur Erstellung des eigenen synthetischen Videos wird nicht im Voraus festgelegt. Durch die laufenden Arbeitsschritte und die damit verbundenen Herausforderungen wird die Vorgehensweise laufend angepasst. Die Entscheidungen für das gewählte Vorgehen werden während des Erstellungsprozesses erläutert. Als Quelle für den Deepfake dienen Videomaterial der Zürcher Hochschule für Angewandte Wissenschaften sowie Filmaufnahmen des Autors. Das Vorgehen bei der Erstellung des eigenen Deepfakes wird durch eine Anleitung aus einem bekannten Deepfake-Forum sowie einem YouTube-Video unterstützt.

Mit einem selbst erstellten Deepfake, der Literatur- sowie Artikelrecherche bei Schweizer Zeitungen werden abschliessend die Erkenntnisse und Resultate diskutiert sowie in ein Fazit formuliert.

## **2 Was ist ein Deepfake?**

Ein Deepfake ist ein von einer KI generierter Inhalt, oft in Form von Videos, der in den Augen eines Menschen authentisch ist (Mirsky & Lee, 2021, S. 1). Das Wort Deepfake ist eine Kontamination und setzt sich aus den Wörtern «Deep Learning» und «Fake» zusammen (Mirsky & Lee, 2021, S. 1; Schick, 2020, S. 46; Verhoeven, 2020, S. 233). Der Begriff umfasst alle durch künstliche Intelligenz (KI) und neuronale Netzwerke visuellen und auditiven Identitätsveränderungen (Mirsky & Lee, 2021, S. 1; Verhoeven, 2020, S. 233). Die häufigste Form von Deepfakes ist der Austausch von Gesichtern in Videos, wobei diese Gesichter durch KI auf die Mimik und Bewegung der Originalperson angepasst werden.

Deepfakes tauchten das erste Mal am 2. November 2017 auf, als ein anonymes «Reddit» mit dem Namen «deepfakes» ein Subreddit-Diskussionsforum eröffnete (Cole, 2017; Schick, 2020, S. 46). Der Redditor widmete sein Forum dem Posten von gefälschten Pornovideos von Hollywood-Schauspielerinnen (Schick, 2020, S. 46). Er erstellte die Videos selbst, indem er KI-Tools und Open-Source-Code verwendete, welche jeder, der sich mit Deep-Learning-Algorithmen auskennt, zusammenstellen konnte (Schick, 2020, S. 47). Durch Google-Bildersuche, Archivfotos und YouTube sammelte «deepfakes» einen Datensatz der Schauspielerin Gal Gadot, mit dem er anschliessend einen KI-Algorithmus

trainierte (Cole, 2017; Whittaker et al., 2020, S. 91). Der Algorithmus lernte, das Gesicht der Schauspielerin Bild für Bild in einen bestehenden Pornofilm zu tauschen, damit es so aussieht, als wäre die echte Gal Gadot die Pornodarstellerin. Obwohl die Verbreitung von gefälschten Promi-Pornos eine der Lieblingsbeschäftigungen des Internets ist, erforderte diese Kreation eine besondere Aufmerksamkeit (Schick, 2020, S. 47). Im Video von «deepfakes» wurden die Emotionen der Pornodarstellerin auf das Gesicht von Gal Gadot projiziert, was ein ganzer Schritt weiter ist, als ein Bild von einer berühmten Person auf einen nackten Pornostar zu photoshopen (Schick, 2020, S. 47). Beim Anschauen des Videos merken die Zuschauer jedoch schnell, dass es in dem synthetischen Inhalt unterschiedliche Pannen gibt. Erste Aufmerksamkeit in der Öffentlichkeit erlangten Deepfakes als die Motherboard-Autorin Samantha Cole rund einen Monat später über das Diskussionsforum auf Reddit berichtete. Der Artikel von Cole sorgte für Aufsehen und innerhalb weniger Wochen schloss Reddit das Forum (Cole, 2018). Der anonyme Reddit-User, welcher das Ganze ins Rollen gebracht hatte, teilte den alles entscheidenden Deepfake-Code öffentlich mit der Community, bevor er verschwand (Cole, 2018; Schick, 2020, S. 50). Seine Pionierarbeit wurde von begeisterten Anhängern aufgegriffen, wobei kurze Zeit später neue Tools und Open-Source-Software im Internet auftauchten, die das Erstellen von Deepfakes vereinfachen (Cole, 2018; Schick, 2020, S. 50). Seit diesem Zeitpunkt explodierte die Erstellung von Deepfakes, die Tools werden immer einfacher und sind leichter verfügbar, parallel dazu sind Deepfakes überzeugender und schwieriger zu entdecken.

Die Grundlage für den rasanten Anstieg an immer authentischeren Deepfakes bilden nebst der Kreativität der Entwickler vor allem die technischen Grundlagen und neue Vorgehensweisen zur Erstellung der gefälschten Inhalte.

### **3 Technologische Grundlagen**

Um zu verstehen, wie Deepfakes erstellt werden, muss zuerst die Technologie verstanden werden, welche Deepfakes möglich macht. Deshalb werden in diesem Kapitel die technischen Grundlagen zur Erstellung eines Deepfakes erläutert sowie die verschiedenen Erzeugungsmethoden aufgezeigt.

Die wichtigste technologische Komponente bei der Erstellung von Deepfakes ist Deep Learning. Diese Technologie ist bekannt für ihre Fähigkeit, komplexe und unstrukturierte Daten darzustellen (Nguyen et al., 2020, S. 1). Dabei bildet Deep Learning eine Untergruppe innerhalb des Bereichs des Machine Learning. Beim Deep Learning wird dem Computer nicht eine endlose Liste zur Lösung eines Problems beigebracht, sondern es wird ein Modell mitgegeben, um Beispiele auszuwerten. Zusätzlich erhält der Computer eine kleine Sammlung von Anweisungen, damit das Modell modifiziert werden kann, wenn Fehler auftreten (Elevenpaths, 2019, S. 8). Mit der Zeit wird erwartet, dass diese Modelle in der Lage sind, Probleme äusserst präzise zu lösen, da das System die Fähigkeit entwickelt, Muster zu erkennen. Obwohl es verschiedene Techniken zur Implementierung von Deep Learning gibt, ist die häufigste die Simulation eines künstlichen neuronalen Netzwerks (Elevenpaths, 2019, S. 8).

Neuronale Netze sind Berechnungssysteme und stellen den Schlüssel zum Erfolg von Machine Learning dar (Heath, 2020; Zucconi, 2018). Die Struktur und Funktionsweise neuronaler Netze ist von den Verbindungen zwischen Neuronen im Gehirn inspiriert (Heath, 2020). Das übliche neuronale Netzwerk besitzt eine Eingabeschicht, eine oder mehrere versteckte Schichten und eine einzelne Ausgabeschicht. Dieser Aufbau wird in Abbildung 1 dargestellt. Jede Schicht hat eine unterschiedliche Anzahl von Neuronen, die in einem dichten Netzwerk miteinander verbunden sind, wodurch Informationen verarbeitet und übertragen werden können (Elevenpaths, 2019, S. 9; Zucconi, 2018). Die Art und Weise, wie die Knoten verbunden sind, bestimmt den Typ des Netzwerkes sowie seine Fähigkeit, die Aufgabe besser auszuführen (Kietzmann et al., 2020, S. 139; Zucconi, 2018). Während des Trainings von neuronalen Netzen werden die Neuronenschichten mit genügend Daten versorgt, um dadurch Muster zu erkennen, zu klassifizieren und zu kategorisieren (Elevenpaths, 2019, S. 9). Ein untrainiertes neuronales Netz verfügt über zufällige Verbindungen zwischen den Einheiten, was zu einem willkürlichen Informationsfluss durch das Netzwerk und damit zu einer zufälligen Ausgabe führt (Kietzmann et al., 2020, S. 139).

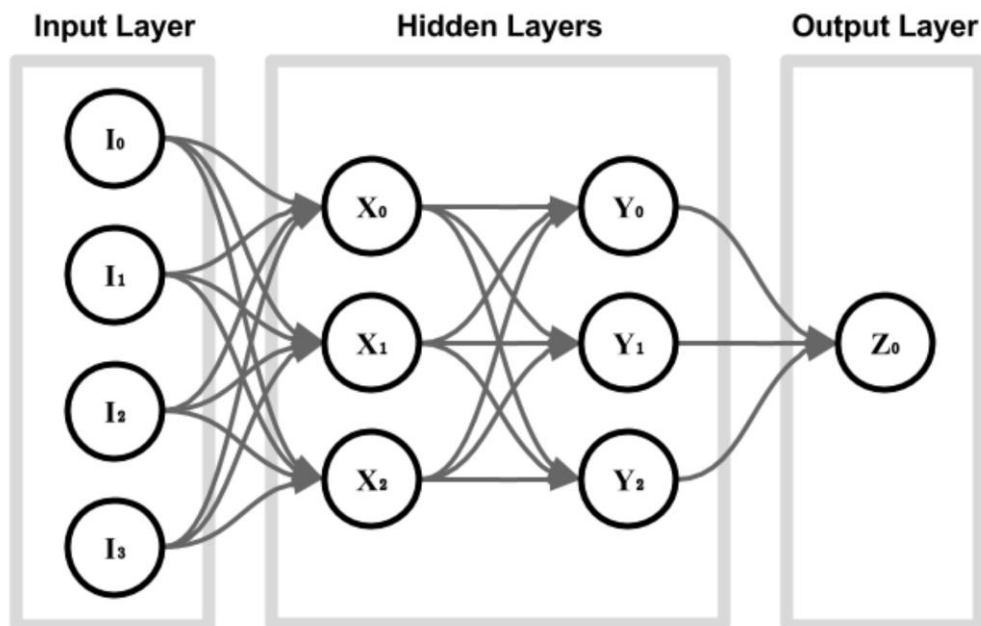


Abbildung 1: Architektur eines neuronalen Netzes (Zucconi, 2018).

Obwohl es eine grosse Anzahl von neuronalen Netzwerken gibt, werden die meisten Deepfakes mit Variationen oder Kombinationen von Autoencoder oder Generative Adversarial Networks (GANs) erstellt. In den folgenden zwei Abschnitten werden deshalb diese beiden Ansätze zur Erstellung von Deepfakes vorgestellt.

### 3.1 Der Autoencoder

Die erste Technologie, die bei der Erstellung von Deepfakes eingesetzt wurde, war der Autoencoder (Katarya & Lal, 2020, S. 480). Um Autoencoder für Deepfakes zu verwenden, werden diese mit einer grossen Anzahl von Gesichtsbildern einer bestimmten Person gefüttert (Whittaker et al., 2020, S. 92). Basierend auf diesem gegebenen Datensatz von Eingabebildern wird ein Autoencoder darauf trainiert, Schlüsselmerkmale eines Gesichts zu erkennen, um anschliessend die Eingabebilder als Ausgabe neu zu erstellen (Kietzmann et al., 2020, S. 139). Der Prozess, zuerst eine kleine Anzahl von Gesichtsmarkmalen in der Eingabe zu erkennen und dann real aussehende Gesichter zu erzeugen, wird in drei Teilbereichen durchgeführt: einem Encoder, einem latenten Raum sowie einem Decoder (Katarya & Lal, 2020, S. 480; Kietzmann et al., 2020, S. 139).

Um die komplexe Aufgabe zu lösen, das gleiche Bild nachzubilden, welches als Eingabewert präsentiert wird, lernen Autoencoder zunächst, abstraktere Gesichtsmerkmale und emotionale Ausdrücke aus dem Eingangsbild zu extrahieren (Katarya & Lal, 2020, S. 480). Der Encoder nimmt zehntausende von Pixeln und komprimiert sie in 300 Messungen, die sich auf bestimmte Gesichtsmerkmale beziehen (Kietzmann et al., 2020, S. 139). Dieses verdichtete Bild wird anschliessend als Eingabe in den latenten Raum gesendet. Der latente Raum ist für das Verstehen und Lernen von Mustern sowie strukturellen Ähnlichkeiten zwischen den Datenpunkten nützlich (Katarya & Lal, 2020, S. 480). Der Decoder dekomprimiert diese Informationen, um eine Ausgabe basierend auf der Repräsentation im latenten Raum zu rekonstruieren. Dabei versucht der Decoder, ein Bild nachzubilden, welches dem Original so weit wie möglich ähnelt (Katarya & Lal, 2020, S. 480; Kietzmann et al., 2020, S. 139). Das Problem besteht jedoch darin, dass der Autoencoder zwar verschiedene Gesichter aus ausgewählten Punkten im latenten Raum erzeugen kann, jedoch der Bildgenerator nicht einfach angewiesen werden kann, daraus ein lächelndes Gesicht zu erzeugen (Kietzmann et al., 2020, S. 140). Die Lösung dieses Problems ist der Trick, der durch Autoencoder erschaffene Deepfakes wie Zauberkunststücke erscheinen lässt.

Gemäss Deshmukh & Wankhade (2021, S. 295) und Katarya & Lal (2020, S. 480) wird für die Erstellung von Deepfakes eine spezielle Konfiguration von zwei Autoencodern angewendet. Damit ein rekonstruiertes Gesicht B aus dem Originalgesicht A erzeugt werden kann, müssen diese zwei Autoencoder trainiert werden. Es ist erforderlich, dass beide denselben Encoder verwenden, da ansonsten die ausgewählten Merkmale im latenten Raum nicht sinnvoll kombiniert werden können. Der Decoder hingegen bleibt personenspezifisch. Der Encoder lernt, allgemeine Merkmale zu verwenden, welche die Gesichter A und B gemeinsam haben. Dadurch entsteht die Möglichkeit, ähnliche Bilder von den zwei verschiedenen Gesichtern an einer ähnlichen Stelle im latenten Raum zu positionieren. Nachdem die Autoencoder trainiert sind, wird der Merkmalsatz von Gesicht A aus dem latenten Raum mit dem Decoder B verknüpft. Dabei wird Gesicht B auf dem Originalgesicht A rekonstruiert. Somit zeigt das Bild denselben emotionalen Ausdruck, dieselbe Kopfhaltung wie das ursprüngliche Gesicht A, jedoch mit dem Gesicht B. In Abbildung 2 wird der beschriebene Ablauf dargestellt.

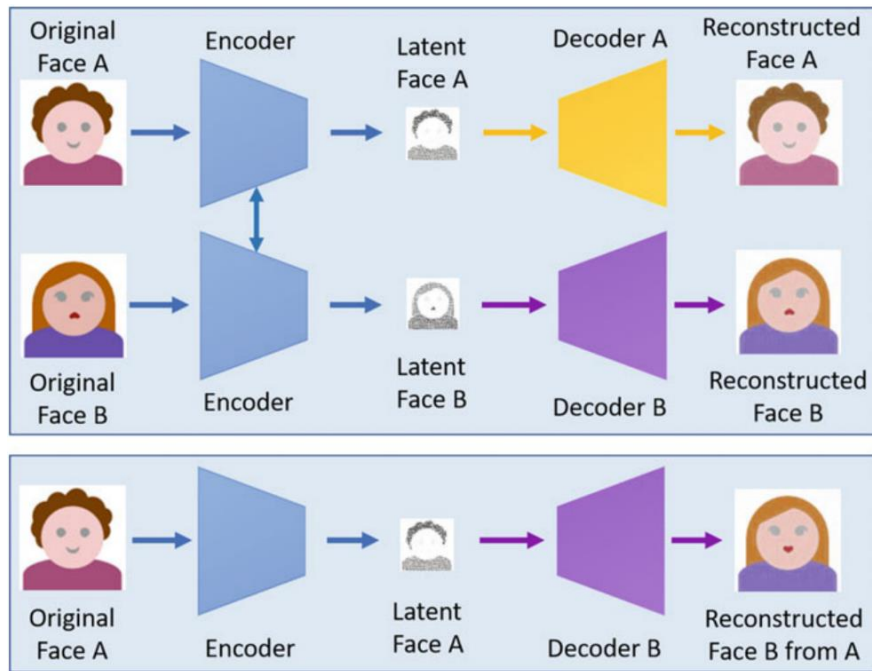


Abbildung 2: Ablauf eines Autoencoders (Deshmukh & Wankhade, 2021, S. 295).

Der Autoencoder lernt somit, ein Modell des Gesichts einer Person zu erstellen. Er ermöglicht ein Bild, welches das Gesicht einer Person zeigt, in eines umzuwandeln, das eine andere Person darstellt. Diese Technologie wurde bereits für den ersten Deepfake verwendet und wird auch heute in einigen Softwares eingesetzt. Mittlerweile hat sich eine zweite Methode für die Erstellung von Deepfakes durchgesetzt.

### 3.2 Generative Adversarial Networks (GANs)

Generative Adversarial Networks wurden vom früheren Google-Forscher Ian Goodfellow erfunden (Pantserev, 2020, S. 40). Sein damaliges Forschungsteam arbeitete an einem Deep-Learning-Projekt, welches das Ziel hatte, mittels künstlicher Intelligenz überzeugend echt aussehende menschliche Gesichter darzustellen (Pantserev, 2020, S. 40). Die frühen Fortschritte im Deep-Learning bedeuteten zu diesem Zeitpunkt, dass Maschinen sehr gut darin waren Daten zu kategorisieren, aber nicht darin, sie zu erzeugen (Schick, 2020, S. 59). In einer Diskussion mit Forschungskollegen an einem Abend im Jahr 2014 kam Goodfellow die Idee, wie ein KI-System es schaffen könnte, echte Gesichter zu erzeugen (Schick, 2020, S. 60). Sein Gedanke war, zwei Deep-Learning-Netzwerke in einem Spiel gegeneinander antreten zu lassen und sich somit gegenseitig zu trainieren (Goodfellow et al., 2014, S. 1; Pantserev, 2020, S. 40). Noch am gleichen Abend setzte sich Goodfellow an die Arbeit und programmierte zwei Deep-Learning-Netzwerke, die



sich in einem gegnerischen Spiel gegenseitig testeten, um menschliche Gesichter zu generieren (Schick, 2020, S. 6). Damit gelang Goodfellow ein unglaublicher Durchbruch, denn innerhalb weniger Stunden kreierte das von ihm aufgebaute System menschliche Gesichter, die besser waren als alles, was KI zuvor in diesem Bereich geschafft hatte (Schick, 2020, S. 61). Obwohl die generierten Gesichter eine geringe Auflösung aufwiesen, zeigten sie deutlich die leistungsstarken Fähigkeiten dieser neuen Art von Trainingsnetzwerken (Whittaker et al., 2020, S. 92).

Für die Generierung von synthetischen Inhalten verfolgen GANs ein bestimmtes Prinzip. Das erste neuronale Netzwerk, der Generator, erzeugt ein neues gefälschtes Video, Bild oder Audio, indem es den heruntergeladenen Datensatz kopiert (Goodfellow et al., 2014, S. 1; Pantserev, 2020, S. 40). Anschliessend werden der Originaldatensatz und der vom ersten neuronalen Netzwerk erstellte gefälschte Inhalt in das zweite neuronale Netzwerk, den Diskriminator, geladen (Pantserev, 2020, S. 40; Whittaker et al., 2020, S. 92). Der Diskriminator wird darauf trainiert, erfolgreich zu erkennen, welcher der Inhalte nicht echt ist (Pantserev, 2020, S. 40; Whittaker et al., 2020, S. 92). Wenn der Diskriminator in der Lage ist, den gefälschten Inhalt zu bestimmen, versucht im Gegenzug der Generator zu lernen, wie der Diskriminator verstanden hat, welcher Inhalt gefälscht ist und nimmt anschliessend entsprechende Korrekturen vor (Pantserev, 2020, S. 40; Whittaker et al., 2020, S. 92). In diesem gegenseitigen Zusammenspiel wollen beide Netzwerke jeweils das andere ausstechen. Sobald die Netze trainiert sind, wird der Diskriminator verworfen und der Generator verwendet, um neue Inhalte zu generieren (Mirsky & Lee, 2021, S. 7; Pantserev, 2020, S. 40).

Seit der Entdeckung von GANs durch Goodfellow wurden im Lauf der Jahre zahlreiche Variationen und Verbesserungen von GANs vorgeschlagen (Mirsky & Lee, 2021, S. 6). Während die ersten Ergebnisse von generierten Gesichtern eine geringe Qualität aufwiesen, sind diese Netzwerke heutzutage in der Lage, hochauflösende Gesichtsbilder von nicht existierenden Personen zu erzeugen (Pantserev, 2020, S. 40). Die schwarz-weißen, körnigen Bilder in Abbildung 3 sind die ersten durch Goodfellow et al. (2014, S. 6) kreierte Gesichtsbilder.



Abbildung 3: Die ersten GAN-Gesichtsbilder (Goodfellow et. al., 2014, S. 6).

Seit der Erfindung von GANs hat sich das Niveau ihrer Ausgaben rasant entwickelt. Die Gesichter in Abbildung 4 zeigen, wie durch GAN erzeugte Bilder heutzutage aussehen. Alle stammen von der Webseite [www.thispersondoesnotexist.com](http://www.thispersondoesnotexist.com). Basierend auf Goodfellow's Forschung erzeugt die Website durch den Einsatz von einem GAN neue menschliche Gesichter. Jedes Mal, wenn der Browser aktualisiert wird, erscheint ein neues gefälschtes KI-Gesicht (This Person Does Not Exist, 2021). Selbst kleine Details wie Falten, Poren und Sommersprossen werden überzeugend dargestellt. Der Vergleich durch die beiden Abbildungen sollte eine Vorstellung davon geben, wie schnell KI voranschreitet. Durch GANs kann KI bereits nahezu perfekte synthetische Bilder erzeugen (Schick, 2020, S. 61). Aufgrund des iterativen und sich gegenseitig aufhebenden Lernprozesses, der von einem GAN verwendet wird, können KI-generierte Medien theoretisch bis zu dem Punkt fortschreiten, an dem sie perfekt werden (Schick, 2020, S. 61). Dadurch wird es mit jeder neuen Iteration schwieriger, zu unterscheiden, ob es sich dabei um einen echten oder gefälschten Inhalt handelt.

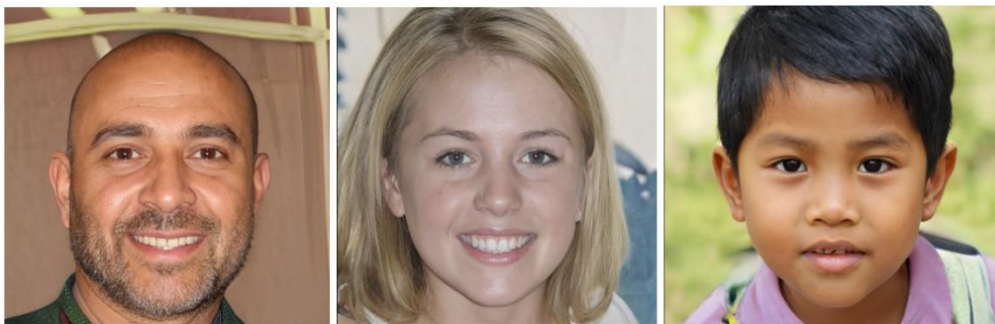


Abbildung 4: Heutige synthetische GAN-Bilder (This Person Does Not Exist, 2021).

### **3.3 Die verschiedenen Erzeugungsmethoden von Deepfakes**

Seit der erste Deepfake 2017 auf Reddit erschienen ist, haben sich die KI basierten Tools zur Erstellung von gefälschten Inhalten rasant verbessert. Dabei sind verschiedene neue Generierungsmethoden von Deepfakes entwickelt worden. Diese Entwicklungen wurden in verschiedenen Arbeiten festgehalten und in drei bis fünf verschiedene Kategorien aufgeteilt (Juefei-Xu et al., 2021, S. 5; Masood et al., 2021, S. 6; Mirsky & Lee, 2021, S. 3; Tolosana et al., 2020, S. 3; Whittaker et al., 2020, S. 91). Durch den Literaturreview können Deepfakes in die folgenden Kategorien unterteilt werden:

- Face-Swap (Gesichtertausch)
- Lip-Syncing (Lippensynchronisation)
- Puppet-mastery (Nachahmung)
- Audio-Deepfakes
- Gesichtssynthese und Merkmalsmanipulation.

Die Unterscheidung in diese fünf Kategorien ist für die Forschung insofern wichtig, weil deshalb neue Methoden zur Entdeckung von Deepfakes entwickelt werden können. Ausserdem wird dadurch ein kleiner Überblick über die Entwicklung von Deepfakes und deren Tools geschaffen. Um aufzuzeigen, wie sich die verschiedenen Erzeugungsmethoden unterscheiden, werden in dem folgenden Abschnitt die oben erwähnten Kategorien erläutert.

#### **3.3.1 Face-Swap**

Bei einem Face-Swap wird das Gesicht der Person im Bild oder Video automatisch durch das Gesicht einer anderen Person ersetzt. Der erste Deepfake, der 2017 auf Reddit hochgeladen wurde, benutzte einen Face-Swap-Ansatz. Die Idee, ein Gesicht auf einem Foto oder Video zu ersetzen, gibt es, seit diese Medien existieren (Maksutov et al., 2020, S. 408). Eines der ersten bekannten Beispiele für eine Bildmanipulation war ein Foto des US-Präsidenten Abraham Lincoln aus dem Jahre 1860. Als Lincoln ermordet wurde, gab es einen Mangel an Bildern des Präsidenten. Um dieses Problem zu beheben, beschloss ein Graveur, eine Fotografie von Lincolns Kopf über eine Gravur des Körpers des Südstaatenpolitikers John C. Calhoun zu legen. Diese Manipulation fiel über ein Jahrhundert lang niemandem auf und wurde erst vor Kurzem entdeckt (Guera & Delp, 2018, S. 1; Maksutov et al., 2020, S. 408).

Bitouk et al. (2008, S. 1) und Masood et al. (2021, S. 6) erklären in ihren Arbeiten, dass traditionelle Face-Swap-Ansätze in der Regel drei Schritte durchlaufen, um einen Face-Swap-Vorgang durchzuführen. Als Erstes erkennen Face-Swap-Tools das Gesicht in den Ursprungsbildern und wählen dann ein Kandidatengesichtsbild aus der Gesichtsbibliothek aus, welches dem Gesicht des Ausgangsbildes in Aussehen und Pose ähnlich ist. In einem nächsten Schritt werden die Augen, die Nase und der Mund des Gesichts ersetzt sowie die Beleuchtung und die Farbe des Kandidatengesichtsbildes an das Aussehen des Ursprungsbildes angepasst. Beide Gesichter werden anschliessend zusammengefügt. Danach wird eine Rangfolge der zusammengeführten Ersetzungskandidaten erstellt, indem ein Übereinstimmungsabstand über den Überlappungsbereich berechnet wird. Dieser Ansatz liefert unter den gegebenen Umständen zwar gute Resultate, besitzt jedoch hauptsächlich zwei Einschränkungen. Erstens wird das Eingabegesicht vollständig durch das Zielgesicht ersetzt, womit die Mimik des Ursprungsbildes verloren geht. Zweitens ist das künstliche Ergebnis sehr starr und das ersetzte Gesicht sieht dadurch unnatürlich aus (Masood et al., 2021, S. 6). Seit den Fortschritten im Bereich des Deep-Learning sind diese Ansätze aufgrund ihrer realistischen Ergebnisse für die Erstellung von gefälschten Medien beliebt geworden.

Nachdem der Reddit-User «deepfakes» in den Medien entlarvt wurde, begannen Forschung und Online-Communities, verbesserte Wege zu finden, um Face-Swapping mit neuronalen Netzwerken durchzuführen (Mirsky & Lee, 2021, S. 24). Die heutigen Face-Swap-Ansätze benutzen häufig Autoencoder und GANs zur Generierung von Deepfakes. Durch die Deep-Learning-Abläufe wurde es einfacher und schneller, Deepfakes mit überzeugenderen Ergebnissen zu erzeugen, wenn ein Gesicht in einem Video durch ein anderes ersetzt, sprich ein Face-Swap eingesetzt wird (Masood et al., 2021, S. 7; Mirsky & Lee, 2021, S. 24). Mirsky & Lee (2021) sowie Masood et al. (2021) haben die verschiedenen Face-Swap-Modelle zusammengetragen und liefern eine aktuelle Übersicht. Beide Publikationen zeigen, wie die Modelle funktionieren und welche Vorteile diese besitzen, um mit dem Ansatz möglichst beeindruckende Deepfakes zu erzeugen.

Um einen einfachen Face-Swap zu erstellen, braucht man heute nicht mehr detailliertes Fachwissen auf dem Gebiet. Mittlerweile gibt es zahlreiche Apps, um einen Face-Swap mit einem Bild oder Video durchzuführen. Auch wenn einige dieser Technologien von

geringer Qualität sind, so gibt es doch auch solche, mit denen die Fälschungen real erscheinen. Ein weiterer Vorteil der Face-Swap-Apps ist deren Schnelligkeit und Verfügbarkeit (Kietzmann et al., 2020, S. 2; Sohrawardi et al., 2019, S. 2613). Eine der fortgeschrittensten Apps, welche man für Android- oder iPhone-Devices herunterladen kann, ist Reface. Abbildung 5 zeigt, wie einfach es ist, per Reface-App mit dem Handy ein Deepfake zu erstellen. Als ersten Schritt muss man einfach ein Selfie mit der Kamera aufnehmen oder kann ein beliebiges Bild aus der Galerie auswählen. Als Beispiel wird ein Portrait des deutschen Stand-Up-Comedian Felix Lobrecht verwendet (Instagram, 2021). Danach kann man zwischen verschiedenen berühmten Filmszenen von Schauspielern und Schauspielerinnen sowie Memes und GIFs auswählen. In diesem Beispiel wird als Ausgangslage das Gesicht des US-Schauspielers Kevin Hart benutzt. Anschliessend braucht die App nur wenige Sekunden, um den Face-Swap mittels GAN-Technologie zu erstellen (Reface, 2021). In der Ausgabe wird schliesslich das Gesicht von Felix Lobrecht mit den Emotionen von Kevin Hart angezeigt.



Abbildung 5: Veranschaulichung des Reface-Prozesses.

Diese Art von Deepfake kann heutzutage sehr einfach und auch schnell erstellt werden und ist die unter Deepfakes am weitesten verbreitete Technik (Agarwal et al., 2020, S. 2814). Jedoch gibt es zwischen den einzelnen Softwares und Apps Unterschiede, was den Aufwand und die Qualität des Deepfake betrifft. Von dieser Art der Manipulation könnten viele verschiedene Branchen profitieren, insbesondere die Filmindustrie. Andererseits besteht jedoch auch eine grosse Gefahr, dass die Technologie für bösartige Zwecke, wie die Erstellung von pornografischen Videos von Prominenten oder Betrügereien, eingesetzt werden kann (Tolosana et al., 2020, S. 3).

### 3.3.2 Lip-Syncing

Beim Ansatz der Lippensynchronisation entsteht ein typischer Deepfake, indem der Mund einer Person im Video so generiert wird, dass er mit einer beliebigen Audioeingabe übereinstimmt (Juefei-Xu et al., 2021, S. 19). Durch die rasanten Fortschritte im Bereich des maschinellen Lernens ist es heutzutage möglich, hochrealistische Audio-, Bild- und Videodateien zu erzeugen, in denen man jeden dazu bringen kann, so gut wie alles zu sagen (Agarwal et al., 2020, S. 2814). Ein wichtiger Aspekt zur Erstellung der visuellen Sprachsynthese ist die Bewegung und das Aussehen des unteren Teils des Mundes und der umgebenden Region. Damit der künstlich erzeugte Inhalt effektiver und natürlicher erscheint, ist es wichtig, die richtigen Lippenbewegungen zusammen mit der Mimik zu erzeugen.

Die ersten Arbeiten im Bereich der Lippensynchronisation demonstrierten Vorgehensweisen, die eine Auswahl von Frames aus einem Video oder einer Transkription zusammen mit den Zielemotionen erforderten, um die Bewegung der Lippen zu synthetisieren (Charles et al., 2016, S. 2; Fan et al., 2015, S. 2). Diese Ansätze sind allerdings auf einen bestimmten emotionalen Zustand beschränkt oder lassen sich nicht gut auf unbekannte Gesichter anwenden. Die neuen Deep-Learning-Modelle sind jedoch in der Lage, die Bewegungen aus Audiomerkmale zu lernen und vorherzusagen (Masood et al., 2021, S. 8). Suwajanakorn et al. (2017, S. 1) präsentierten einen Ansatz für die Erstellung eines fotorealistischen, lippensynchronen Videos, indem das Zielvideo und ein beliebiger Audio-clip als Eingabe verwendet werden. In ihrer Forschungsarbeit wurde Barack Obama aufgrund ausreichend verfügbaren Online-Videomaterials als Fallbeispiel verwendet. Das eingesetzte Modell wandelt zuerst die Audioeingabe in eine zeitlich variierende, spärliche Mundkontur um. Basierend auf dieser Mundform wird anschliessend eine fotorealistische Mundtextur erzeugt, die in den Mundbereich des Zielvideos mit Barack Obama komprimiert wird. Bevor die beiden Komponenten endgültig zusammengesetzt werden, müssen die Mundtextur-Abfolge und das Zielvideo abgeglichen und zeitlich abgestimmt werden. Dadurch erscheint die Kopfbewegung möglichst natürlich und passt zur eingegebenen Sprache. Diese Vorgehensweise bewahrt die Details der Lippen sowie Zähne und reproduziert zeitlich variierende Falten und Grübchen um den Mund und das Kinn. In Abbildung 6 wird der Ablauf visuell dargestellt. Suwajanakorn et al. (2017, S. 2) konnten mit ihrem Modell eine glaubwürdige Nachstellung von Obama produzieren, welche die Qualität der bisherigen Arbeiten übertrifft.

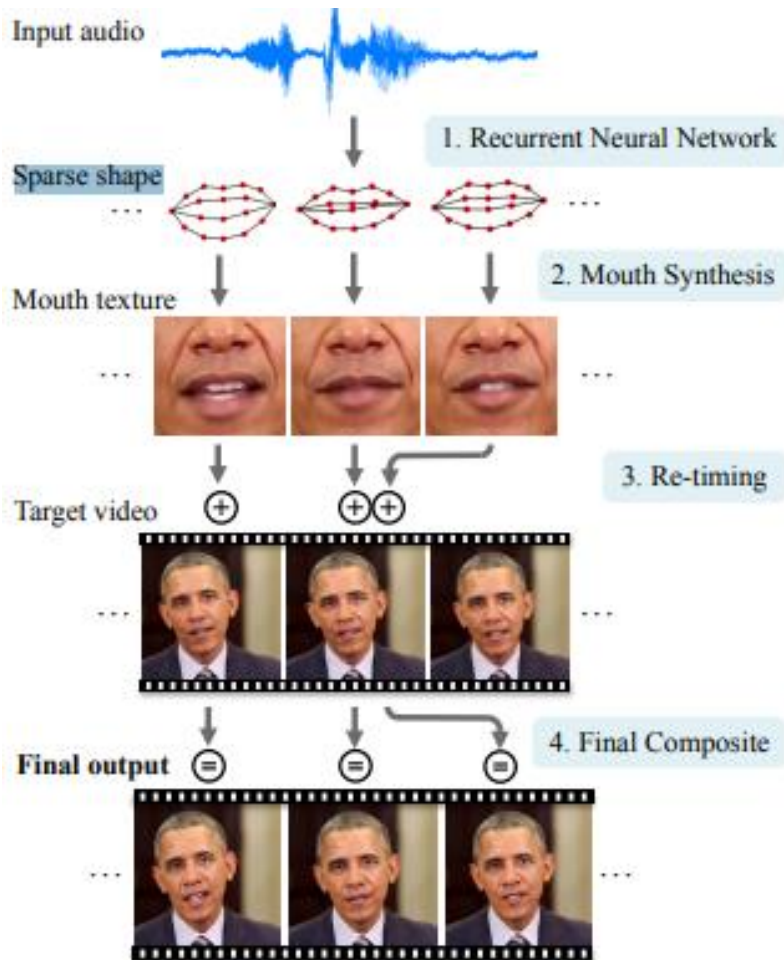


Abbildung 6: Erstellung Lippensynchronisation (Suwajanakorn et al., 2017, S. 2).

Heutzutage gibt es einige Ansätze zur Erstellung von Deepfakes mittels Lippensynchronisation (Masood et al., 2021, S. 9; Mirsky & Lee, 2021, S. 10). Die Umwandlung von Audio in Video ist nicht nur aus rein wissenschaftlicher Sicht interessant, sondern hat auch eine Reihe wichtiger praktischer Anwendungen. Zum Beispiel können audiogesteuerte, fotorealistische digitale Charaktere in Filmen oder Videospiele erstellt werden. Der Ansatz wird auch bei hörgeschädigten Personen angewendet, indem diese von den Lippen eines Videos ablesen, welches mit echtem Ton erstellt wurde (Masood et al., 2021, S. 8; Suwajanakorn et al., 2017, S. 1). Für die Erstellung eines Deepfakes durch Lippensynchronisation braucht es für das Training viele Stunden von Filmmaterial. Ausserdem gibt es zurzeit keine öffentlich zugänglichen Datenbanken in Bezug auf diesen Ansatz. Deshalb erfolgt die Forschung zu diesem Thema in der Regel durch die Synthese eigener Daten unter Verwendung von öffentlich verfügbarem Material.

Darum können Deepfakes unter Verwendung dieses Ansatzes aktuell nur bei prominenten Zielen wie Schauspielern, CEOs und politischen Führungskräften durchgeführt werden (Mirsky & Lee, 2021, S. 30; Tolosana et al., 2020, S. 17)

### **3.3.3 Puppet Master**

Puppet Master (Gesichtsnachstellung) ist eine weitere Variante von Deepfakes, die es einer Person ermöglicht, ein Video mit ihren eigenen Bewegungen und Ausdrücken zu manipulieren (Greengard, 2019, S. 17). Bei diesem Vorgehen zur Erstellung eines Deepfakes wird die Mimik einer Zielperson so manipuliert, dass die Gesichtsgesten, Augen- und Kopfbewegungen von einer Ausgangsperson in das Video übertragen werden (Masood et al., 2021, S. 9). Somit wird die Gesichtsnachstellung verwendet, um eine Zielperson so erscheinen zu lassen, dass sie wie die Quelle handelt und spricht (Sohrawardi et al., 2019, S. 2613). Wie Mirsky & Lee (2021, S. 11) in ihrer Arbeit festhalten, gab es Mimik-Veränderungen schon einige Jahre vor der Popularisierung von Deepfakes. In dem Verfahren von Blanz et al. (2003, S. 1) wird eine Methode zur fotorealistischen Animation beliebiger Gesichter in einem Einzelbild oder einem Video vorgestellt. Dabei formten die Forscher die Modelle durch einen 3D-Scan der Gesichter. Im Jahr 2005 wurde gezeigt, wie durch einen Modelltransfer- und Adaptionalgorithmus eine neuartige Person mit nur einem kleinen Videoausschnitt animiert werden kann (Chang & Ezzat, 2005, S. 143). Thies et al. (2016, S. 1) präsentierten 2015 die erste Methode zur Übertragung von Gesichtsausdrücken in Echtzeit von einem Schauspieler auf eine Zielperson. Die Forschenden verwendeten dabei einen RGB-Sensor (Red, Green, Blue), um das 3D-Modell eines Ausgangs- und eines Zielschauspielers zu verfolgen und zu rekonstruieren.

In den heutigen Modellen werden GANs häufig, aufgrund ihrer Fähigkeit fotorealistische Bilder zu erzeugen, erfolgreich für die Nachstellung von Gesichtern eingesetzt (Masood et al., 2021, S. 10). Kim et al. (2018, S. 12) stellen in ihrer Arbeit den ersten Ansatz, der vollständige fotorealistische Videoporträts des Oberkörpers einer Zielperson, einschliesslich realistischer Kleidung und Haare sowie eines konsistenten Szenenhintergrunds erzeugt, vor. Den Kern ihrer Vorgehensweise bildet ein Conditional Generative Adversarial Network (cGAN), welches speziell auf die Video-Portraitsynthese zugeschnitten ist.



Anstatt nur den Gesichtsausdruck der Zielidentität zu modifizieren, wird bei der Reanimation des Portraitvideos auch das Ändern der Kopfhaltung, des Augenaufschlags und des Blinzeln berücksichtigt. Dabei erreichten sie mit nur einer Minute Training eine komplette Gesichtsnachstellung.

Ebenfalls 2018 präsentieren Wu et al. (2018, S. 107) ein neuartiges lernbasiertes Framework für die Gesichtsnachstellung. Ihre vorgeschlagene Methode, ReenactGAN, basiert auf drei Komponenten: Einem Encoder, um ein Eingabegesicht in einen latenten Raum zu kodieren, einem zielspezifischen Transformator, um einen beliebigen Quellrandraum an den eines bestimmten Ziels anzupassen, und einem zielspezifischen Decoder, der den latenten Raum in das Zielgesicht dekodiert. Dadurch ist die Methode in der Lage, Gesichtsbewegungen und -ausdrücke von der Videoeingabe einer beliebigen Person auf das Video einer Zielperson zu übertragen.

Das Konzept von Thies et al. (2016) war der Start für die erweiterten Ansätze der Gesichtsnachstellung. In den nächsten Jahren werden auch fotorealistische Ganzkörper-Nachstellungs-Videos realisierbar sein, bei denen die Mimik der Zielperson zusammen mit den Bewegungen manipuliert werden, um realistische Deepfakes zu erzeugen (Masood et al., 2021, S. 10). Die Videos, die mit den oben genannten Techniken generiert werden, können mit gefälschtem Audio zusammengeführt werden. Diese Mischung aus Gesichts- und Audiomanipulation machen es gleichzeitig sehr schwierig, die generierten Inhalte von den echten zu unterscheiden, weshalb Puppet Master eine der gefährlichsten Arten von Deepfakes darstellt (Maksutov et al., 2020, S. 408).

### **3.3.4 Audio-Deepfakes**

Jüngste Fortschritte bei KI-synthetisierten Algorithmen zur Sprachsynthese und zum Klonen von Stimmen haben das Potenzial gezeigt, realistische gefälschte Stimmen zu erzeugen, die von echter Sprache kaum zu unterscheiden sind (Masood et al., 2021, S. 15). KI-synthetisierte Audiomanipulation ist eine Art von Deepfake, welche die Stimme einer Person imitiert und diese so darstellen kann, dass die Person etwas ausspricht, das sie in Wahrheit nie gesagt hat (Masood et al., 2021, S. 15). Während Audio-Deepfakes auf den gleichen Prinzipien der Berechnung mit neuronalen Netzwerken beruhen wie visuelle Deepfakes, ist die Herangehensweise an die Verarbeitung von Stimmmaterial aufgrund der Ausgangslage anders (Hauser, 2021).

Damit ein qualitativ hochwertiger Audio-Deepfake erstellt werden kann, braucht es klare Aufnahmen des Sprechers, möglichst ohne Unterbrechungen, Hintergrund- oder Störgeräusche (Hauser, 2021). Eine der grössten Innovationen beim Klonen der Stimmen ist die allgemeine Reduzierung der Menge an Rohdaten, da in der Vergangenheit die Systeme Dutzende oder sogar Hunderte von Stunden an Audiodaten benötigten (Johnson, 2020). Wie die Publikationen von Arik et al. (2018, S. 9) und Lorenzo-Trueba et al. (2018, S. 7) zeigen, kann synthetische Sprache mit bereits wenigen Audio-Beispielen generiert werden, sodass diese wie die Zielperson klingt. Sprachsynthese bezieht sich dabei auf eine Technologie, die Sprache aus einer gegebenen Eingabe generiert (Masood et al., 2021, S. 15). Die Komplexität unserer gesprochenen Sprache und die menschlichen Fehler beim Sprechen, wie Variation des Tempos, Pausen und Füllwörter, stellen eines der grössten Probleme für künstliche Stimmen dar (Stucke, 2020). Aus diesen Gründen hören sich die Ergebnisse oftmals zu perfekt an (Stucke, 2020). Wichtige Entwicklungen im Bereich der Stimm- und Sprachsynthese sind Softwares wie WaveNet, Tacotron und DeepVoice3, die realistisch klingende synthetische Sprache als Texteingaben erzeugen können, um ein verbessertes Interaktionserlebnis zwischen Mensch und Maschine zu ermöglichen (Masood et al., 2021, S. 15). Dabei könnten die gefälschten Stimmen dieser Sprachsynthesemodelle mit Audiosynthese-Programmen wie Audacity kombiniert werden (Masood et al., 2021, S. 15). Die Software zur Audioverarbeitung kann verwendet werden, um die verschiedenen Teile der ursprünglichen und synthetisierten Audios zu kombinieren, wodurch leistungsfähigere und authentischere Audios entstehen (Masood et al., 2021, S. 15).

Durch den enormen Fortschritt der Technologien im Bereich synthetisch generierte Audio gibt es in der Wirtschaft eine enorme Nachfrage (Johnson, 2020). Besonders im Sektor der Spielindustrie konnten durch die neuen Ansätze grosse Erfolge gefeiert werden (Johnson, 2020). Studios haben die Möglichkeit bekommen, die Stimme eines Schauspielers zu klonen und die neuen Ansätze zu verwenden, damit die Charaktere alles in Echtzeit sagen können (Johnson, 2020). Abgesehen davon wurden auch Weiterentwicklungen in den Bereichen der Chatbots, KI-Assistenten, beim Kundensupport sowie in der Werbung erzielt (Johnson, 2020; Masood et al., 2021, S. 15). In der Medizin verspricht die moderne Nachahmung von Stimmen noch etwas Besseres als die bereits existierenden Ansätze. Die Sprachsynthese-Firma CereProc gab 2008 dem mittlerweile verstorbenen Filmkritiker Roger Ebert in einem erstmaligen Verfahren seine Stimme zurück, nachdem

der Krebs sie ihm genommen hatte (Johnson, 2020). In den letzten Jahren haben eine Reihe von Unternehmen mit der ALS Association im Rahmen des Projekts «Revoice» zusammengearbeitet (Johnson, 2020; Project Revoice, 2021). Das Ziel des Projektes ist, dass Menschen, die an ALS (amyotrophe Lateralsklerose) leiden und ihre Stimme durch die Krankheit verloren haben, eine synthetische Stimme als Ersatz erhalten (Project Revoice, 2021). Neben den positiven Eigenschaften von synthetisch generierten Stimmen gibt es jedoch auch eine grosse Gefahr von böswilligem Missbrauch. In Kombination mit visuellen Deepfakes lässt sich damit potenziell eine vollständige Imitation einer Person erstellen, wodurch Deepfakes noch gefährlicher werden (Hauser, 2021).

### **3.3.5 Gesichtssynthese und Merkmalmanipulation**

Die Bearbeitung von Gesichtern in digitalen Bildern sowie die Videosynthese werden schon seit mehreren Jahren, bevor Deepfakes berühmt wurden, erforscht (Zhang et al., 2020, S. 67). In letzter Zeit werden diese Techniken jedoch auch vermehrt zur Erstellung von Deepfakes genutzt (Masood et al., 2021, S. 11). Die Manipulation von Gesichtern wird in den Arbeiten von Juefei-Xu et al. (2021, S. 7) und Masood et al. (2020, S. 11) in zwei Kategorien unterteilt: Gesichtssynthese und Bearbeitung von Gesichtsattributen. Bei der Erzeugung von Gesichtern geht es um die Synthese von fotorealistischen Bildern eines menschlichen Gesichtes, das im echten Leben nicht existiert. Das Ziel ist es, möglichst echte nichtexistierende Gesichtsbilder zu erzeugen. Im Gegensatz dazu handelt es sich bei der Bearbeitung von Gesichtsattributen um die Veränderung des Gesichtsaussehens eines vorhandenen Musters, indem bestimmte Attribute überarbeitet werden, während die irrelevanten Bereiche unverändert bleiben. Die Bearbeitung von Gesichtsattributen umfasst unter anderem die Retusche der Haut, die Veränderung der Haarfarbe und sogar einige komplexere Attribute wie die Änderung des Alters oder des Geschlechts.

Die enorme Entwicklung im Bereich der GANs hat dazu geführt, dass diese Modelle weit verbreitete Werkzeuge für die Bildsynthese und -bearbeitung sind (Masood et al., 2021, S. 11). Seit der Entdeckung von GANs präsentierten Radford et al. (2015) DCGAN als ersten Ansatz für eine vollständige Bildsynthese. Zwei Jahre später gab es eine explosionsartige Zunahme an vertiefter Forschung zu GANs. Die bahnbrechende Arbeit war der WassersteinGAN (WGAN) von Arjovsky et al. (2017, S. 9). Der Algorithmus stellte dabei eine Alternative zum traditionellen GAN-Training dar (Arjovsky et al., 2017, S. 16). Mit dem neuen Modell wurde die Stabilität des Lernens verbessert und einige Probleme, wie die Einhaltung des Gleichgewichts zwischen Generator und Diskriminator, gelöst

(Arjovsky et al., 2017, S. 16). Basierend auf dem WGAN sind weitere Ansätze wie der StyleGAN entstanden, welcher eine sehr hohe Auflösung der generierten Bilder mit feinen Details erreichte (Karras et al., 2018, S. 1). Die Auflösung der generierten Bilder beträgt mindestens 1024x1024 Pixel und mittlerweile sind die Bilder so detailliert, dass es schwierig ist, diese zwischen echten und gefälschten Bildern zu unterscheiden (Karras et al., 2018, S. 1).

Basierend auf der GAN-Technologie wurden in den letzten Jahren auch einige Ansätze vorgeschlagen, um Gesichtsattribute wie Hautfarbe, Frisur, Alter oder Gesichtsausdruck zu bearbeiten (Masood et al., 2021, S. 12). In Abbildung 7 ist ersichtlich, wie das Originalbild durch die verschiedenen Gesichtsm Manipulationen verändert und mittels GAN-Technologie generiert wird (Masood et al., 2021, S. 12). Der IcGAN von Perarnau et al. (2016, S. 8) war der erste Versuch, Gesichtsattribute zu manipulieren. Bei dieser Vorgehensweise bildet der Encoder das eingegebene Gesichtsbild im latenten Raum und einem Attributbearbeitungsvektor ab, wobei der veränderte Attributvektor als Bedingung mitgegeben wird. Dadurch kommt es zu Informationsverlusten und die ursprüngliche Gesichtside ntität im neu erzeugten Bild wird verändert. Um das Problem der Skalierbarkeit und Robustheit bei den bisherigen Ansätzen zu beheben, stellten Choi et al. (2018, S. 1) StarGAN vor. Die einheitliche Modellarchitektur von StarGAN ermöglicht das gleichzeitige Training mehrerer Datensätze mit unterschiedlichen Domänen innerhalb eines einzigen Netzwerkes. Durch StarGAN werden damit, im Vergleich zu bereits existierenden Modellen, Bilder in überlegener Qualität generiert.



Abbildung 7: Kategorien von Gesichtsm Manipulationen (Masood et al., 2021, S. 12).

Die Gesichtssynthese hat bereits eine Vielzahl von nützlichen Anwendungen, wie die automatische Charaktererstellung für Videospiele und die 3D-Gesichtsmodellierung, ermöglicht (Tolosana et al., 2020, S. 3). Für die Gesichtsbearbeitung nutzen Menschen kommerziell verfügbare KI-basierte mobile Anwendungen wie FaceApp, welche das Aussehen eines Eingabebildes automatisch verändern (FaceApp, 2021). Die Benutzer können dabei mit verschiedenen KI-Filtern, Hintergründen, Effekten und anderen Werkzeugen in nur einem Klick das Eingabebild verändern (FaceApp, 2021). Diese Technologien können aber auch für schädliche Anwendungen genutzt werden, indem sehr realistische Fake-Profilen in den sozialen Netzwerken erstellt werden, um Fehlinformationen zu verbreiten (Masood et al., 2021, S. 11).

### **3.4 Herausforderungen**

Seit 2017 haben sich Deepfake-Generierungsmethoden rasant entwickelt (Juefei-Xu et al., 2021, S. 11). Bei den fünf erwähnten Hauptkategorien hat die Qualität der generierten Bilder dazu geführt, dass es für menschliche Augen extrem schwer zu erkennen ist, ob es sich um echte oder gefälschte Inhalte handelt (Juefei-Xu et al., 2021, S. 11). Trotz der rasanten Entwicklung der Generierungsmethoden, gibt es nach Masood et al. (2020, S. 18) und Juefei-Xu et al. (2021, S. 26) bei der Erstellung von Deepfakes mehrere offene Herausforderungen, wovon einige in diesem Abschnitt aufgeführt werden.

**Posenschwankungen und Abstand zur Kamera:** Die bestehenden Deepfake-Techniken erzeugen gute Ergebnisse für die frontale Gesichtsansicht. Die Qualität der erzeugten Inhalte nimmt jedoch deutlich ab, wenn die Person vor der Kamera wegschaut. Ebenso führt eine Zunahme des Abstands zur Kamera zu einer qualitativ schlechteren Gesichtssynthese.

**Beleuchtungsbedingungen:** Die aktuellen Ansätze produzieren die gefälschten Inhalte in einer kontrollierten Umgebung mit unveränderten Lichtverhältnissen. Ein abrupter Wechsel der Beleuchtungsbedingungen hat Farbinconsistenzen und seltsame Artefakte in den resultierenden Videos zur Folge.

**Fehlende Steigerung der Auflösung:** Seit die erste GAN-Methode zur Erzeugung hochauflösender Bilder (1024x1024) vorgestellt wurde, haben die neuen Methoden zur Synthese von gefälschten Inhalten keine höhere Auflösung entwickelt.

**Begrenzte Eigenschaften der Methoden zur Gesichtsmanipulation:** Aktuell können Methoden zur Attributmanipulation nur die Eigenschaften verändern, die durch den Trainingsatz vorgegeben sind. Daher sind diese Methoden eingeschränkt und es müsste eine Methode geben, die unabhängig von den Eigenschaften des Trainingssets ist.

**Synthetisches Audio:** Obwohl sich die Qualität beim synthetischen Audio durch die Forschungsarbeiten verbessert hat, fehlt den audiobasierten Deepfakes die Menschlichkeit. Zu den Schwierigkeiten gehören natürliche Emotionen, Pausen, Atemlosigkeit und das Tempo, mit welchem die Zielperson spricht.

**Okklusionen:** Eine der grössten Herausforderungen bei Deepfakes ist das Auftreten von Okklusionen, wenn der Gesichtsbereich der Quelle und des Ziels durch eine Hand, Haare, Brille oder andere Gegenstände verdeckt ist.

Die Erkenntnisse über die Herausforderungen für die Erstellung von Deepfakes zeigen, in welchen Bereichen in Zukunft die Forschung ihre Schwerpunkte haben könnte. Eine deutliche Verbesserung der Deepfake Generierungsmethoden wird wiederum die Entwicklung der Deepfake Entdeckungsmethoden vorantreiben. Mit den Verbesserungen im Bereich der Deepfake-Generierungsmethoden, werden sich die verschiedenen Kategorien anfangen zu vermischen, wodurch die Unterscheidung vom blossen Auge für die Menschen immer schwieriger wird. Nebst den verschiedenen Erzeugungsmethoden von Deepfakes ist es jedoch für die Forschung auch wichtig zu wissen, wie fest die Verbreitung der gefälschten Inhalte vorangeschritten ist.

## 4 Verbreitung von Deepfakes

Um zu verstehen, wie weit Deepfakes in der Gesellschaft vorangeschritten sind, wird die Verbreitung der gefälschten Inhalte analysiert. Dabei wird dieses Kapitel in zwei Sektionen unterteilt. Im ersten Teil wird die weltweite Verbreitung von Deepfakes behandelt, während im zweiten Teil eine Analyse der medialen Berichte über Deepfakes in der Schweiz durchgeführt wird. Anhand dieser beiden Themen soll die weltweite Verbreitung von Deepfakes sowie das Bewusstsein über die Technologie in der schweizerischen Gesellschaft aufgezeigt werden.

## 4.1 Weltweite Landschaft von Deepfakes

Im Bereich der Verbreitung von Deepfakes existieren bis heute keine wissenschaftlichen Arbeiten. Beim Literaturreview wurde deutlich, dass die Forscher und Forscherinnen hauptsächlich neue Methoden zur Erkennung von Deepfakes oder neue Generierungsmodelle entwickeln. Abgesehen von Arbeiten, die den aktuellen Stand der bisherigen veröffentlichten Methoden von Deepfakes zusammenfassen, gibt es keine wissenschaftliche Arbeit, welche sich mit der aktuellen Landschaft von Deepfakes auseinandersetzt. Die einzigen Recherchen, die zu dem Thema durchgeführt wurden, stammen von den zwei Firmen Sentinel AI und Sensitiv AI. Die Reports dieser beiden Unternehmen dienen als Grundlage für diesen Abschnitt.

Nachdem Ende 2017 das Phänomen Deepfakes durch Zeitungsartikel bekannt wurde, tauchten im Jahr 2018 die ersten Deepfake-Videos im Internet auf. Sensitive AI hat durch ihre Recherche herausgefunden, dass die Zahlen seit 2018 von veröffentlichten Deepfakes exponentiell in die Höhe steigen (Cavalli, 2021). Dabei verdoppeln sich die Zahlen etwa alle sechs Monate und sind von 7'964 Deepfakes im Dezember 2018 auf 85'047 im Dezember 2020 angestiegen (Cavalli, 2021). Dieser rasante Anstieg wird durch die zunehmende Kommerzialisierung von Tools und Diensten unterstützt, welche die Barriere für Anfänger von der Erstellung gefälschter Inhalte senken (Ajder et al., 2019, S. 5). Der enorme Zuwachs der online gestellten Deepfakes ist in Abbildung 8 ersichtlich. Bei der Sammlung der Daten berücksichtigte Sensitive AI die Plattformen YouTube, Vimeo, LiveLeak, Dailymotion sowie Pornoplattformen (Ajder et al., 2019, S. 23).

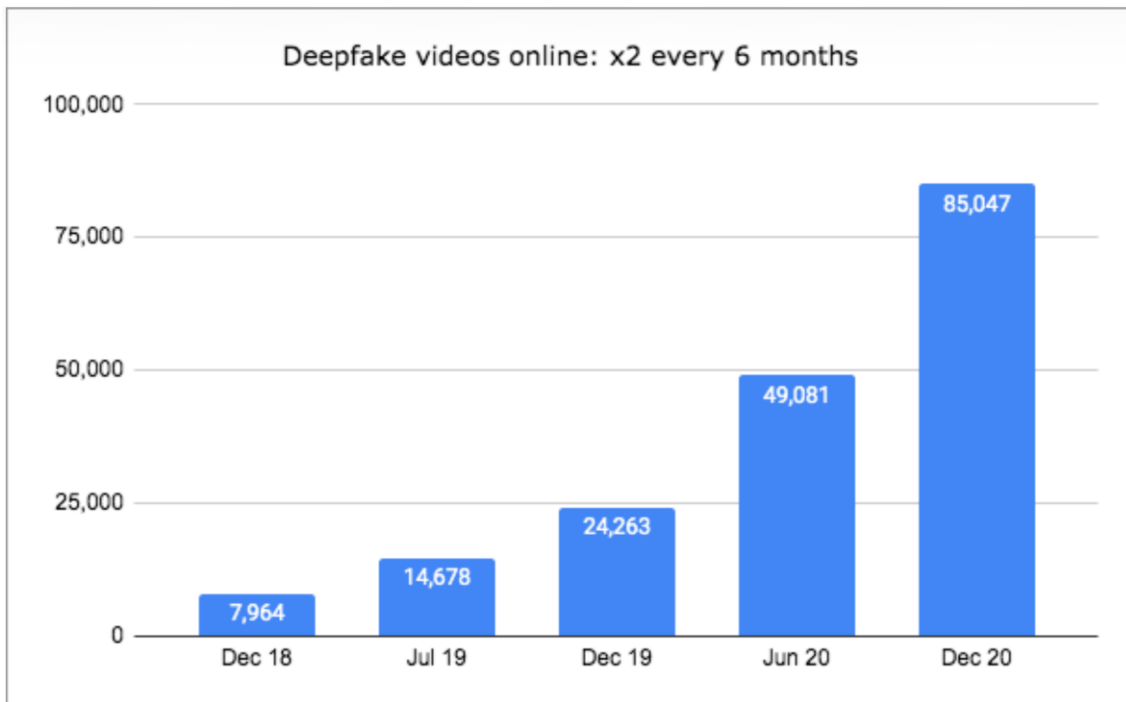


Abbildung 8: Online gestellte Deepfakes (Cavalli, 2021).

Während Sensitive AI in ihrem Report bereits ein beeindruckendes Wachstum feststellen, präsentierte die Firma Sentinel AI für Juni 2020 eine noch grössere Anzahl an Deepfakes, welche in Darstellung 9 ersichtlich ist (Tammekänd et al., 2020, S. 7). Der Unterschied in der Anzahl entdeckter Deepfakes lässt sich durch die Datenquellen erklären. Sentinel AI berücksichtigte für die Sammlung der Daten folgende Plattformen: YouTube, Facebook, TikTok, Instagram, Twitter, Yoku, Iqiyi, Tencent Video, Bilibilli, Nicovideo, Vimeo, Dailymotion sowie über 30 Pornowebsites (Tammekänd et al., 2020, S. 81). Dabei wurden 43 Prozent der nicht-pornografischen Deepfakes auf Twitter verbreitet, gefolgt von 32 Prozent auf YouTube (Tammekänd et al., 2020, S. 8).

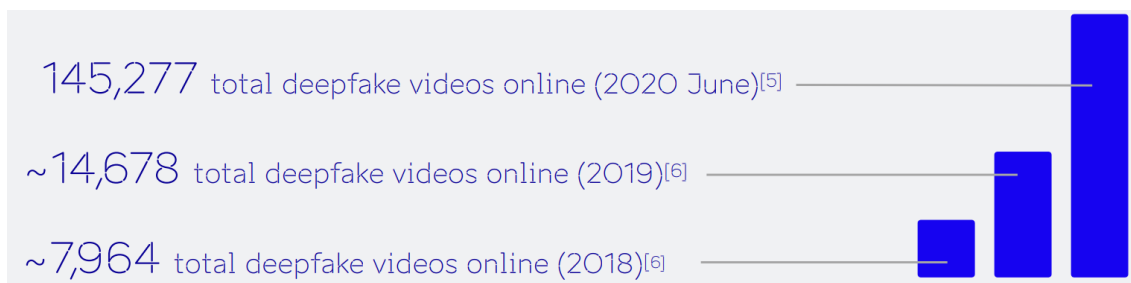


Abbildung 9: Entdeckter Deepfakes (Tammekänd et al., 2020, S. 7).



Wie in Kapitel zwei bereits beschrieben wurde, sind GANs, aufgrund der flexiblen Anwendung und den realistischen Ausgaben, derzeit die beliebteste Technik zur Erzeugung von gefälschten Inhalten. Grafik 10 von Ajder et al. (2019, S. 9) zeigt die Messungen von Sensitive AI über den Forschungsoutput zu GANs seit ihrer Erfindung im Jahr 2014. Aus dem Ergebnis ergibt sich ein indirekter Hinweis darauf, wie rasch sich die Qualität von GANs verbessert und wie schnell neue Möglichkeiten zur Erstellung von Deepfakes entwickelt werden.

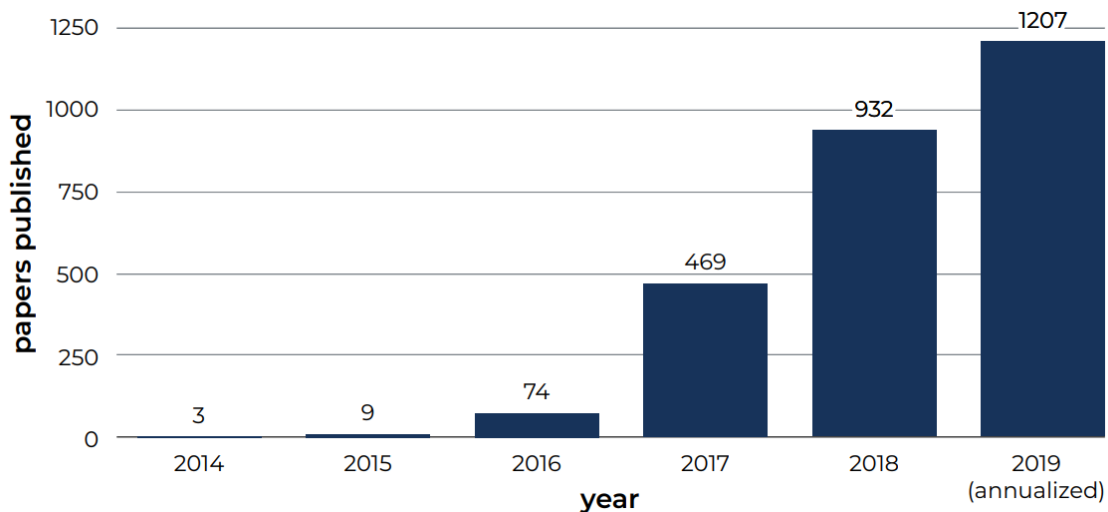


Abbildung 10: Akademische Arbeiten über GANs (Ajder et al., 2019, S. 9).

Tammekänd et al. (2020, S. 81) haben in einem Untersuchungszeitraum von November 2017 bis Mai 2020 herausgefunden, dass es sich bei 81 Prozent der nicht pornografischen Deepfakes um Face-Swaps handelt. Rund 18 Prozent sind Gesichtsnachahmungen und nur 1 Prozent wurde als Ganzkörper-Deepfakes eingestuft. Ausserdem ergab die Datenauswertung, dass 63 Prozent der nicht-pornografischen Deepfakes für Entertainment erstellt wurden. Dabei sind diese gefälschten Inhalte als Belustigung für den Zuschauer gedacht. An zweiter Stelle mit je 14 Prozent sind politische sowie sachbezogene Deepfakes.

Um die Demografie aufzuzeigen, analysierte Sensitive AI, Deepfake-Videos der Top fünf Deepfake-Pornografie-Webseiten und von 14 Video-Kanälen zur Erstellung von gefälschten Inhalten auf YouTube (Ajder et al., 2019, S. 8). Dabei achteten die Forschenden auf das Geschlecht, die Nationalität und den Beruf der Personen, welche im Video dargestellt wurden (Ajder et al., 2019, S. 8). Die Ergebnisse sind in den Darstellungen 11 bis 13 abgebildet.

Beim Geschlecht zeigt sich eines der grössten Probleme von böswilligen Deepfakes. Die Datenerhebung von Ajder et al. (2020, S. 8) ergab nämlich, dass Deepfake-Pornografie ausschliesslich auf Frauen abzielt. Im Gegensatz dazu enthielten die gefälschten Videos auf YouTube mit 61 Prozent mehrheitlich männliche Protagonisten.

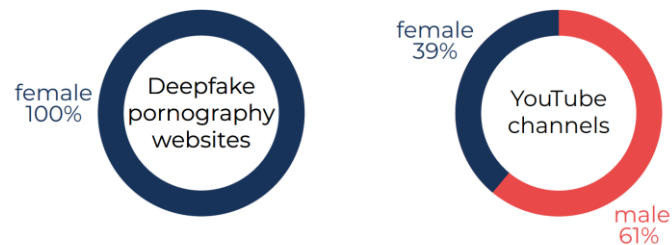


Abbildung 11: Unterschied der Geschlechter (Ajder et al., 2020, S. 8).

Über 90 Prozent der Deepfake-Videos auf YouTube enthielten westliche Probanden. In pornografischen Deepfakes sind jedoch in fast einem Drittel der Videos nicht-westliche Subjekte vorhanden, wobei südkoreanische K-Pop-Sängerinnen rund einen Viertel der Zielpersonen ausmachten. Aus dieser Auswertung kann man indirekt ableiten, dass Deepfake-Pornografie ein zunehmend globales Phänomen wird und sich die neue Technologie in mehreren Teilen der Welt ausbreitet.

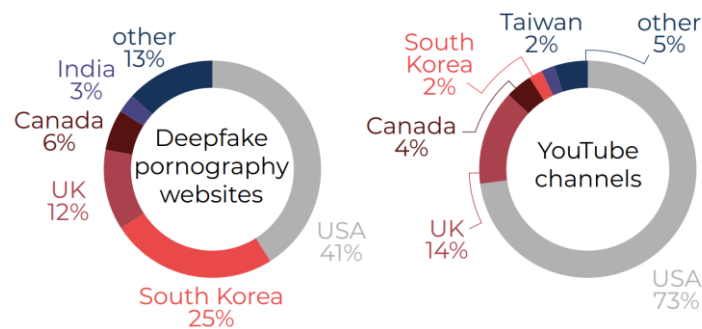


Abbildung 12: Ein globales Phänomen (Ajder et al., 2020, S. 8).

Bis auf 1 Prozent waren alle Frauen, die in den pornografischen Videos auftauchen, Schauspieler- und Musikerinnen aus der Unterhaltungsbranche. Die Protagonisten in den YouTube Videos kommen aus einer breiteren Berufsgruppe, darunter sind nebst Personen aus der Unterhaltungsbranche vor allem Politiker und Unternehmensvertreter.

Diese Zusammenstellung lässt sich dadurch erklären, dass heutzutage für einen überzeugenden Deepfake genug Videomaterial benötigt wird. Personen, welche oft in der Öffentlichkeit stehen, sind daher ideale Ziele zur Erstellung von Deepfakes, weil die Daten für jeden zugänglich sind.

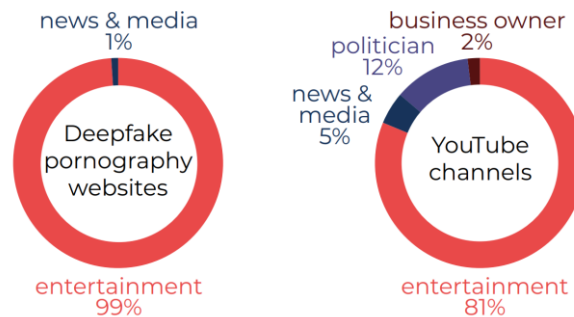


Abbildung 13: Protagonisten und deren Berufe (Ajder et al., 2020, S. 8).

Die zwei Berichte der Firmen Sensitive AI und Sentinel AI haben wichtige Schlussfolgerungen zum aktuellen Stand von Deepfakes geliefert. Durch die wachsende Anzahl von Deepfake-Erstellungstechnologien und -werkzeugen ist es möglich, Deepfakes immer einfacher zu erstellen. Dazu kommt eine wachsende Anzahl von Apps und Online-Diensten, welche die Technologie zunehmend kommerzialisieren. Die Online-Präsenz von Deepfakes verzeichnete in den letzten Jahren eine massive Zunahme, wobei die überwiegende Mehrheit der gefälschten Videos pornografische Inhalte enthält (Ajder et al., 2020, S. 15; Kietzmann et al., 2020, S. 142). In Abbildung 12 wurde dabei gezeigt, dass es sich bei Deepfake-Pornografie zunehmend auch um ein globales Problem handelt, wobei ausschliesslich Frauen betroffen sind. Die Geschwindigkeit der Entwicklungen rund um Deepfakes bedeutet jedoch, dass sich die Landschaft ständig verändert, womit es in Zukunft wichtig ist, in diesem Bereich vertiefter zu forschen. Ebenfalls fehlen Studien über die Verbreitung und das Bewusstsein von Deepfakes in der Gesellschaft einzelner Länder.

## 4.2 Medialer Umfang über Deepfakes in der Schweiz

Damit eine gefährliche Technologie wie Deepfakes von der Bevölkerung richtig eingeschätzt werden kann, braucht es in erster Linie ein allgemeines Verständnis in der Gesellschaft. Für ein solches Verständnis können zum Beispiel Medien wie Zeitungen, Radio oder Fernsehsendungen genutzt werden. Aus diesem Grund wird zur Untersuchung in diesem Kapitel über das Bewusstsein von Deepfakes in der Schweizer Öffentlichkeit in verschiedenen schweizerischen Zeitungen nach dem Begriff Deepfake gesucht. Dies soll einen Überblick verschaffen, wie oft seit 2017 über Deepfakes in den Medien berichtet

wurde und die Annahme bekräftigen, dass in der Schweiz eine grosse Mehrheit der Bevölkerung die Bedeutung des Ausdrucks nicht kennt. Bei der Suche wurden dabei Zeitungen von drei Landessprachen berücksichtigt: Deutsch, Französisch und Italienisch. Ebenso wurden zwei Online-News-Portale in die Recherche miteinbezogen. In den daraus resultierenden Tabellen sind die Titel der Artikel, das Datum der Veröffentlichung sowie der Name der Zeitung eingetragen. Es wurden alle Publikationen berücksichtigt, bei welchen das Wort Deepfake mindestens einmal vorkommt.

Tabelle 1 präsentiert die Ergebnisse der Artikelsuche bei den deutschsprachigen Zeitungen. In den vier ausgewählten Zeitungen wurden insgesamt 48 Artikel mit dem Begriff entdeckt. Im Zeitraum von Dezember 2017 bis April 2021 verfasste der Tages-Anzeiger mit 21 Beiträgen am meisten Berichte. 2019 wurden mit insgesamt 22 Ergebnissen am meisten Publikationen veröffentlicht, gefolgt von 11 Publikationen im Jahre 2020. Um eine möglichst grosse Fläche abdecken zu können, wurden auch andere Zeitungen wie die Basler Zeitung, der Bund und die Südostschweiz berücksichtigt. Die Basler Zeitung und der Bund gehören zur selben Mediengruppe wie der Tages-Anzeiger, weshalb bei der Suche die genau gleichen Artikel erschienen sind. Aus diesem Grund wurden diese Zeitungen in Tabelle 1 nicht aufgeführt. Die Graubündner Zeitung Südostschweiz hat bei der Suche keine Ergebnisse angezeigt.

Zeitung:	Artikel:	Datum:
20 Minuten	Unternehmen photoshoppt dich in Sexfilme	29.08.2018
	Dieses Fake-Geständnis von Zuckerberg geht viral	12.06.2019
	Betrüger gibt sich als Chef aus und klaut Millionen	09.07.2019
	Diese App schockiert gerade das Internet	03.09.2019
	Forscher tricksen Gesichtserkennung aus	29.10.2019
	Falsche Nacktbilder von mehr als 100'000 Frauen veröffentlicht	21.10.2020
	Dieser falsche Tom Cruise hat Millionen Views auf TikTok	01.03.2021
Tages-Anzeiger	Fälscher macht «Wonder Woman» zur Porno-Darstellerin	14.12.2017
	Menschen müssen aufhören, ihren Augen zu trauen	12.02.2018
	Erschreckend echt	09.03.2018
	Diese Technologien können Angst machen	21.05.2018
	Pentagon erklärt Fake-Videos den Krieg	13.06.2018
	Deepfakes: So leicht lassen sich Gesichter in Videos fälschen	27.10.2018
	Jetzt kommt ein Gegengift gegen Fake-Videos	05.01.2019
	Facebook verschärft seine Livestream-Regeln	15.05.2019
	Die Mona Lisa lernt sprechen	26.05.2019
	Millionen sahen ein Fake-Video, das Pelosi blamiert	28.05.2019
	Gegen Deepfakes hilft nur eines: Faktencheck	13.06.2019
	Videos sind glaubwürdig – das war einmal	14.07.2019
	Youtube kassiert mehr Werbegelder als die gesamte TV-Konkurrenz	06.02.2020
	Twitter kennzeichnet von Trump geteiltes Video als «manipuliert»	09.03.2020
	Neue App macht Fake-Videos massentauglich	21.07.2020
	Bastien Girods Meinungsumschwung: «SUVs sind grossartig»	31.12.2020
	«Deepfakes wurden durch Pornografie bekannt»	28.07.2020
	Queen Elizabeth: «Sie sind nicht allein»	25.12.2020
	Der gefälschte Tom Cruise ist der bisher beste Deepfake	01.03.2021
	Wer ist schuld am Fälschungsskandal?	25.03.2021

	«Facebook und Tiktok sind für mich Hochrisiko-Technologie»	22.04.2021
St. Galler Tagblatt	Nazi Göring krächzt nicht mehr: Die Tondokumente der Nürnberger Prozesse sind nun komplett digitalisiert	14.09.2019
	Nicht alle Kreuzlinger Behördenmitglieder und Wirtschaftsvertreter gehen in die Ferien	11.01.2019
	Wenn aus Popstars Pornostars gemacht werden	08.09.2020
	Berset tanzt oben ohne mit «KK Sexy»: Ostschweizer Band Dachs lanciert erstes Deep-Fake-Musikvideo	22.03.2021
Neue Zürcher Zeitung	Künstliche Intelligenz macht Trump zum Dummkopf	19.04.2018
	Wie sich die EU und die Nato auf hybride Angriffe vorbereiten	18.12.2018
	Deepfake: Auch was falsch ist, ist irgendwie war. Nur anders	19.12.2018
	«Global Risk»-Briefing: Trübe Aussichten für das bevölkerungsreichste Land Afrikas	14.02.2019
	Deepfakes – so erkennt man manipulierte Videos	11.03.2019
	Deepfakes: Kann ich überhaupt noch glauben, was ich sehe?	11.03.2019
	Ein Video zeigt eine betrunkene Nancy Pelosi – und führt uns vor Augen, was mit Deepfakes heute alles möglich ist	25.05.2019
	Künstliche Intelligenz bringt Mona Lisa zum Reden	28.05.2019
	Warum Facebook ein gefälschtes Video von CEO Mark Zuckerberg nicht löscht	12.06.2019
	Ein Blinzeln lässt die Fälscher auffliegen	04.07.2019
	Die jüngere Generation wird den Fake-News den Garaus machen	23.07.2019
	Irgendetwas ist immer wahr – auch wenn Deepfakes widerlegt werden, werfen sie die Dividende des Lügners ab	09.08.2019
	Um Betrug bei der Präsidentenwahl vorzubeugen, verbietet Kalifornien Deepfakes	14.10.2019
	Dieser Deepfake-Profi wollte, dass alle Videos fälschen können – heute warnt er davor, dass die Technik ausser Kontrolle gerät	30.01.2020
	Israelische Spionage-Software Pegasus soll beim Khashoggi-Mord eine Rolle gespielt haben	11.03.2020
Coronavirus und Falschmeldungen: Die «Infodemie» ist real	15.04.2020	

Tabelle 1: Zeitungsartikel der deutschsprachigen Zeitungen.

Für den französischen Sprachraum wurde bei den Zeitungen Le Temps, 24 heures und ArcInfo nach dem Begriff recherchiert. Diese Zeitungen stammen aus drei von vier rein französisch sprechenden Kantonen und hatten 2020 zusammengerechnet eine Auflage von etwa 90'000 Printausgaben (WEMF AG, 2020, S. 2). Insgesamt wurden bei den drei Zeitungen 25 Artikel gefunden. Der erste Bericht erschien am 30. Januar 2018 und handelte über das erste Deepfake-Video aus dem Jahre 2017. Wie auch im deutschsprachigen Zeitraum sind 2019 die meisten Publikationen erschienen, gefolgt vom Jahr 2020. Die Artikel und deren Erscheinungsdatum sind Tabelle 2 zu entnehmen.

Zeitung:	Artikel:	Datum:
Le Temps	Les ravages du «fake porn», à la portée de tous	04.02.2018
	Facebook est responsable face aux deepfakes	30.06.2019
	L'information vue comme bien public	28.09.2019
	Désormais, il faudra cesser de croire vos yeux	12.10.2019
	Deepfakes : des faux plus vrais que nature	16.10.2019
	En 2069, quand l'EPFL sera centenaire	19.12.2019
	Samsung présente ses «Neons», êtres virtuels intelligents et «dotés d'émotions»	07.01.2020
	Entre l'intelligence artificielle et l'éthique, une relation tendue	28.01.2020
	Jean-Marc Rickli, pour qui le monde est une jungle implacable, avec des menaces, partout	18.03.2020
	«Lisez et découvrez la merveilleuse semaine des médias au collège des Co-teaux !»	19.12.2020
	24 heures	Le « deepfake », le porno des stars
Un Obama plus vrai que nature insulte Trump		30.04.2018
Mark Zuckerberg piégé par une vidéo truquée		12.06.2019
« Nous n'écoutons pas dans votre microphone »		25.06.2019
L'application qui déshabillait les femmes supprimée		29.06.2019

	Les fausses vidéos menacent la présidentielle	23.09.2019
	Voici le visage moyen du parlementaire suisse	30.09.2019
	Samsung dévoile un avatar humain très réaliste	07.01.2020
	Twitter : ses usager augmentent, l'action s'envole	06.02.2020
	L'intelligence artificielle crée des visages 100% réalistes	15.01.2021
	Dieudonné condamné pour injure publique envers une magistrate	14.04.2021
ArcInfo	Mark Zuckerberg piégé sur Instagram : la video truquée restera en ligne	12.06.2019
	Smartphones : « DeepNude », l'application qui déshabillait les femmes, supprimée	29.06.2019
	« Trumperies sur la marchandise », l'ar du temps du Luc-Olivier Erard	30.10.2020
	Deep Nostalgia : pourquoi l'app qui donne vie aux photos anciennes doit être utilisée avec prudence	08.03.2021

Tabelle 2: Artikel der französischsprachigen Schweiz.

Die italienische Schweiz deckt nur 8.1 Prozent der Schweizer Bevölkerung ab (Tourismus Schweiz, 2021), weshalb die Auswahl von grösseren Zeitungen eher gering ist. Die Resultate der Suche sind in Tabelle 3 abgebildet. Dabei wurde der Corriere del Ticino ausgewählt, weil er im Tessin die grösste Auflage besitzt (WEMF AG, 2020, S. 9). Bei der Zeitung mit der zweitgrössten Auflage, laRegione, kann jedoch nur mit einem Abo nach Publikationen gesucht werden, weshalb die italienische Version der Boulevardzeitung 20 minuten ausgewählt wurde, welche 2011 in Zusammenarbeit durch Tamedia und laRegione gegründet wurde. Die Suche bei den beiden Zeitungen ergab lediglich vier Treffer, wobei der erste Artikel über Deepfakes im Januar 2019 erschienen ist.

Zeitung:	Artikel:	Datum:
Corriere del Ticino	Attrici e showgirl nel porno, tra sosia e falsi perfetti	03.01.2019
	Musica, creatività e intelligenze artificiali	02.01.2020
20 minuti	Deepfake: quei video-bufala che fanno paura	10.10.2019
	Facebook mette al bando I deepfake	07.01.2020

Tabelle 3: Artikel der italienischsprachigen Zeitungen.

Die Schweiz zählt zu einem der Länder mit der grössten Mediendichte weltweit. Diese ausgeprägte Medienlandschaft wird zunehmend durch digitale Messenger und Social-Media-Dienste abgelöst (Lüscher, 2020). Digitale Kommunikationsmittel besitzen einen immer grösser werdenden Anteil am täglichen Medienkonsum (Lüscher, 2020). Aus diesem Grund wurden auch zwei Nachrichtenportale nach dem Begriff durchsucht. Die Ergebnisse sind in Tabelle 4 ersichtlich und zeigen eine grosse Anzahl von Berichten. Insgesamt wurden auf den beiden Nachrichtenportalen Watson und Nau 38 Publikationen gefunden. Im Unterschied zu den Zeitungen wurden bei den Nachrichtenportalen im Jahr 2020 die meisten Artikel veröffentlicht. Viele der öffentlichen Artikel sind kürzer als bei den Zeitungsportalen.

Zeitung:	Artikel:	Datum:
Watson	Stell dir vor, dein Sexvideo taucht im Internet auf – dabei hast du gar nie eines gedreht	03.02.2018
	Dreht Obama durch? Dieses Video zeigt eine der grössten Gefahren für Demokratien	18.04.2018
	«Absolut krass und krank»: Schweizer Internet-User werden mit Kinderpornos erpresst	21.08.2018
	Merkel hatte recht! Das Internet «ist für uns alle Neuland»	06.12.2018
	Diese Deepfakes musst du gesehen haben	20.02.2019
	Deepfake-Technik erweckt Mona Lisa zum Leben – und bald dein Profilbild?	24.05.2019
	Ein Deepfake von Mark Zuckerberg geistert rum – darum wird's nicht gelöscht	12.06.2019
	Die DeepNude-App zieht Frauen aus – ohne Einwilligung der Betroffenen	02.07.2019
	Deepfake gerät ausser Kontrolle – wie dieser falsche Tom Cruise zeigt	13.08.2019
	10 Deepfake-Videos, bei denen du die Augen reibst. Vor Staunen und vor Lachen	15.08.2019
	Warum hat Jim Carrey plötzlich Brüste?	30.08.2019
	Deepfake-App aus China zeigt, wie einfach sich Videos manipulieren lassen	04.09.2019
	Facebook verbietet Deepfake-Videos – das musst du wissen	08.01.2020
	Dieses Deepfake-Videos lassen dir das Blut in den Adern gefrieren	05.09.2020
	Telegram-Bot verbreitet gefälschte Nacktbilder von Frauen	23.10.2020
	Deepfake-Video zeigt, wie Sean Connery als Gandalf ausgesehen hätte	10.12.2020
	Der gefälschte Tom Cruise: Darum sollten uns Deepfakes beunruhigen	26.02.2021
	Dieses Online-Tool erweckt alte Fotos zum Leben	02.03.2021
Weil Deepfakes machen so einfach ist, haben wir es ausprobiert (sorry, Herr Berset)	15.03.2021	
Nau.ch	Mit einfacher Software – Wonder Womans Gesicht auf Porno-Film	13.12.2017
	«Verstörend» - SRF steckt Büssi in einen Schwulen Porno	09.11.2018
	Instagram will Zuckerberg keine Vorzugsbehandlung nach Fake-Video einräumen	12.06.2019
	Wie Deepfake ausser Kontrolle gerät	14.08.2019
	USA beauftragen Militär mit der Bekämpfung von Fake News	01.09.2019
	Deepfake: Neue App Zao begeistert China	03.09.2019
	Eine neue Ära? Samsung stellt künstliche Menschen vor	05.01.2020
	Samsung will künstliche Avatare echt wirken lassen	07.01.2020
	Facebook verbannt vor US-Wahlen sogenannte Deepfake-Videos	07.01.2020
	Neon: Samsung-Tochterfirma will künstliche Menschen erschaffen haben	09.01.2020
	Von Klimaaktivisten geprägte Endung «for Future» Anglizismus des Jahres	28.01.2020
	Twitter will manipulierte Videos löschen	05.02.2020
	Twitter will «Deepfakes» und manipulierte Inhalte stärker bekämpfen	05.02.2020
	Twitter kennzeichnet von Trump geteiltes Video als «manipuliert»	09.03.2020
	Microsoft veröffentlicht Tool zur Erkennung von Deepfakes	02.09.2020
	Skandal um Telegram: Bots erstellen auf Wunsch nacktbilder	22.10.2020
	Queen-Deepfake: Sender erbost Briten auf Weihnachten mit Satire-Video	24.12.2020
TikTok-Deepfake von Tom Cruise hat Tausende Follower	02.03.2021	
US-Mutter belästigte Cheerleader mit gefälschten Nacktfotos	16.03.2021	

Tabelle 4: Artikel der Online-Nachrichtenportale.

Insgesamt wurden bei den elf ausgewählten Zeitungen und Nachrichtenportalen im Zeitraum von Dezember 2017 bis April 2021 115 Artikel gefunden. Dabei wurden die meisten Berichte im Jahr 2019 veröffentlicht. In diesem Jahr wurden insgesamt 48 Artikel publiziert, wobei am meisten über den Deepfake von Facebook-Gründer Mark Zuckerberg, neue Apps sowie die DeepNude-Applikation geschrieben wurde. Im Jahre 2020 ging die Berichterstattung über Deepfakes bereits wieder zurück und betrug noch 35 Publikatio-

nen. Die Ergebnisse sind in Tabelle fünf dargestellt. Am meisten über Deepfakes berichtete der Tages-Anzeiger mit insgesamt 21 veröffentlichten Artikeln. Zusätzlich enthält der Tages-Anzeiger die ausführlichsten Berichterstattungen zum Thema Deepfakes, wobei auch die Technologie genauer erläutert wird.

In den meisten Zeitungen und Nachrichtenportalen handelt es sich bei den gefundenen Artikeln um kürzere Berichte über neue Gefahren der technologischen Entwicklungen, wobei Deepfakes kurz erwähnt und mit wenigen Worten beschrieben werden. Daher thematisieren die wenigsten Publikationen das Thema Deepfakes als Hauptinhalt. Oftmals wird über generierte Deepfakes geschrieben, welche für grosses Aufsehen in der Öffentlichkeit gesorgt haben, wie der Deepfake von Obama oder der neuste Deepfake von Tom Cruise. Bei den Online-Nachrichtenportalen kommen auch sehr kurze Berichte vor, bei welchen der Hauptinhalt ein Video ist, wo die Besten oder lustigsten Deepfakes gezeigt werden. Einen Unterschied gibt es auch zwischen den einzelnen Regionen. Während die deutsch- und französischsprachigen Zeitungen mehr über Deepfakes berichten, gibt es im italienischsprachigen Raum wenig Artikel über das Thema. Dies hängt jedoch auch mit der Schweizer Demographie zusammen, da die italienischsprachige Bevölkerung deutlich kleiner ist als die deutsch- und französischsprachige. Ebenso ist dies auch eine Frage der finanziellen Mittel, über welche eine Zeitung verfügt.

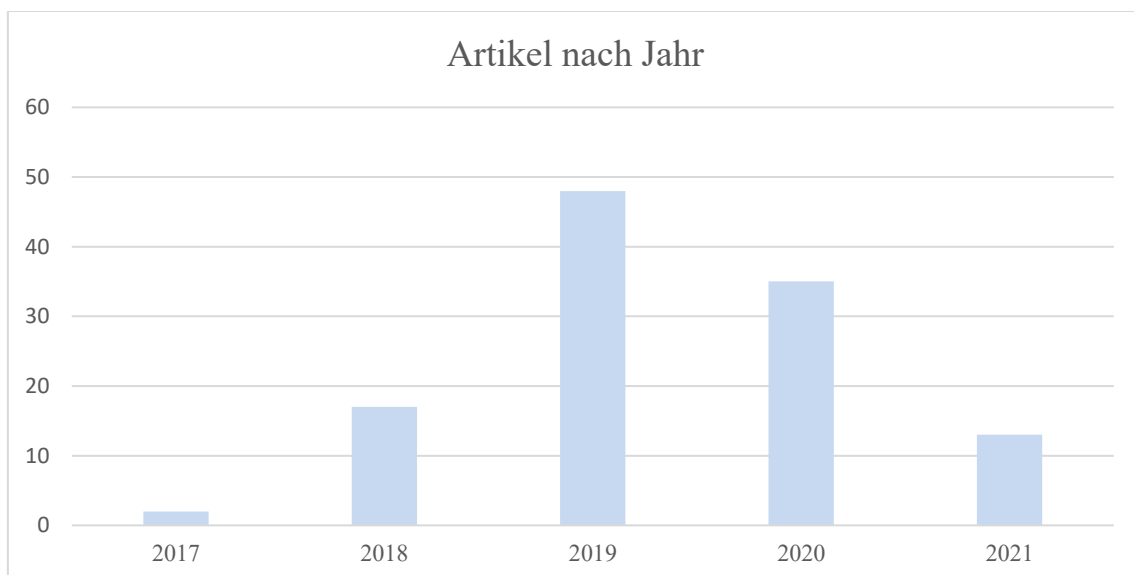


Tabelle 5: Alle Artikel nach Jahr.



Durch die Recherche und die Sichtung der Artikel kann man die indirekte Annahme treffen, dass weiterhin ein grosser Teil der Schweizer Bevölkerung nicht weiss, was ein Deepfake ist. Obwohl in einigen Zeitungen ausführlich über das Thema berichtet wurde, handelt es sich bei den meisten Publikationen um sehr kurze Ausführungen des Begriffs. Eine Studie der Online-Sicherheitsfirma iProov aus dem Jahre 2020 bestärkt diese Annahme. Das Unternehmen führte bei ihren Kunden in Grossbritannien und den Vereinigten Staaten von Amerika eine Umfrage zum Thema Deepfake durch (iProov, 2020, S. 2). Lediglich 13 Prozent der Befragten gaben an, dass sie wissen, was ein Deepfake ist, wobei über 57 Prozent der Kunden noch nie den Begriff Deepfake gehört haben (iProov, 2020, S. 7). In der Schweiz existiert bisher keine Studie über die gesellschaftliche Bekanntheit von Deepfakes. Aufgrund der recherchierten Artikel in den ausgewählten Zeitungen und Nachrichtenportalen kann jedoch davon ausgegangen werden, dass in der Schweiz die Bekanntheit von Deepfakes gleich gering ist wie in Grossbritannien und den Vereinigten Staaten von Amerika. Für eine Bekräftigung der Annahme müsste jedoch zusätzlich noch eine Umfrage durchgeführt werden, welche die verschiedenen Altersgruppen und Regionen der Schweiz umfasst. Ebenso sollen die Medien vermehrt und ausführlicher über Deepfakes berichten, damit die Bevölkerung über Nutzen und Gefahren der Technologie aufgeklärt wird.

## **5 Nutzen und Bedrohungen von Deepfakes**

Mittlerweile steht fest, dass die Technologien, die gefälschte Inhalte generieren können, sehr mächtig sind. Aktuell werden Deepfakes hauptsächlich genutzt, um andere zu täuschen und potenziell auszubeuten (Kietzmann et al., 2020, S. 7). Oftmals ist es so, dass neue Technologien das Gute und das Schlechte im Menschen fördern und die Menschheit sich dadurch gleichzeitig vorwärts und rückwärts bewegt (Kietzmann et al., 2020, S. 7). Deepfakes sind dabei keine Ausnahme und haben wie andere neue Technologien eine helle und eine dunkle Seite (Kietzmann et al., 2020, S. 7; Schick, 2020, S. 9). Die Fortschritte im Bereich der Technologie von Deepfakes bieten zahlreiche Nutzen und Bedrohungen, derer sich die Menschheit bewusst sein sollte (Whittaker et al., 2020, S. 94). Aus diesen Gründen werden in diesem Kapitel die grössten Nutzen und Bedrohungen erläutert, die Deepfakes für Einzelpersonen, Organisationen und Regierungen darstellen. Dabei handelt es sich um einen Themenbereich, der sich in den nächsten Jahren sicherlich ändern und erweitern wird, wenn Deepfakes eine grössere Anwendung in der Gesellschaft finden.

## **5.1 Nutzen durch Deepfakes**

Obwohl Deepfakes oftmals für böswillige Zwecke mit schlechten Absichten verwendet werden, hat diese Technologie auch einige positive Anwendungen in verschiedenen Bereichen (Caporusso, 2021, S. 236). Während zu Recht Fragen über die Konsequenzen durch Deepfakes gestellt werden, ist es wichtig, dass man das Bewusstsein über die positiven Einsatzmöglichkeiten der Technologie bewahrt. Wissenschaftler haben in letzter Zeit vermehrt damit begonnen, die potenziellen Vorteile von Deepfakes zu berücksichtigen und zu erforschen (Caporusso, 2021, S. 236). In Anbetracht der Neuheit dieser Technologie ist es relevant, immer wieder neue Ideen vorzuschlagen, welche die vorteilhaften Aspekte für deren Einsatz hervorheben, während die Debatte über die zukünftige Ausrichtung noch im Gange ist (Caporusso, 2021, S. 236). Deshalb werden einige bedeutende Anwendungen von Deepfakes in den folgenden Abschnitten diskutiert.

### **5.1.1 Bildung**

Schulen und Lehrer verwenden seit längerer Zeit Medien, Audio- und Videomaterial für den Unterricht im Klassenzimmer (Jaiman, 2020). Dabei kann die Deepfake-Technologie Pädagogen helfen, innovative Lektionen zu vermitteln, die wesentlich ansprechender sind als traditionelle visuelle und mediale Formate (Jaiman, 2020). Nach Chesney und Citron (2018, S. 1769) wird es mit Deepfakes möglich sein, Videos von historischen Persönlichkeiten zu produzieren, welche direkt zu den Studenten sprechen und dadurch einer ansonsten unattraktiven Vorlesung neues Leben verleihen. Die Technologie öffnet die Tür zu einer billigen und zugänglichen Produktion von Videoinhalten, die bestehende Filme oder Sendungen verändern, um einen pädagogischen Punkt zu veranschaulichen. Zum Beispiel könnte eine Szene aus einem Kriegsfilm so verändert werden, dass ein Kommandant und sein Rechtsberater über die Anwendung der Kriegsgesetze diskutieren, während im Original der Dialog nichts damit zu tun hat. Die Szene könnte immer wieder mit Änderungen des Dialogs wiederholt werden, die den Änderungen des betrachteten hypothetischen Szenarios entsprechen. Der pädagogische Wert von Deepfakes kann sogar über den Unterricht im Klassenzimmer hinaus gehen. Es gibt die Vorstellung, dass Deepfakes eingesetzt werden, um Aufklärungskampagnen von gemeinnützigen Organisationen zu unterstützen.

Im Jahr 2019 präsentierte die Firma Udacity den ersten Ansatz, wie KI benutzt werden kann, um Unterrichtsvideos automatisch zu produzieren (B.-H. Kim & Ganapathi, 2019, S. 1). Die Produktion von Inhalten für Massive-Open-Online-Course (MOOC)-Plattformen wie Udacity ist akademisch lohnend und lukrativ, jedoch vor allem im Bereich der Videoproduktion enorm zeitaufwändig (B.-H. Kim & Ganapathi, 2019, S. 1). Aus diesem Grund untersuchten Kim und Ganapathi (2019, S. 8) ein neues Framework für maschinelles Lernen, welches automatisch Vorlesungsvideos ausschliesslich aus der Audiosprache generiert. Die vorgestellte KI produziert somit ein synthetisches Vorlesungsvideo mit vollständiger Pose des Dozenten anhand der Audioeingabe. Die Ergebnisse von Kim und Ganapathi (2019, S. 8) können für die akademische Welt in Bezug auf die aufwändige Videoproduktion einen Wendepunkt darstellen und sollen neue Entwicklungen von Deepfake und Deep-Learning-Technologien für die kommerzielle Produktion von Videoinhalten beschleunigen.

### **5.1.2 Kunst**

Deepfake-Technologien werden bereits genutzt, um neue Kunstwerke zu schaffen, das Publikum zu involvieren und einzigartige Erfahrungen zu ermöglichen. Das Dal-í-Museum in St. Petersburg, Florida, setzte Deepfake-Technologie ein, um dem 1989 verstorbenen spanischen Künstler Leben einzuflössen (Mahmud & Sharmin, 2020, S. 14). Unter der Verwendung von Archivmaterial aus Interviews wurden über 6000 Einzelbilder ausgewertet, worauf der KI-Algorithmus 1000 Stunden auf Dalís Gesicht trainiert wurde (Lee, 2019). Anschliessend wurde seine Mimik auf einen Schauspieler mit Dalís Körperproportionen übertragen und Zitate aus seinen Interviews mit einem Sprecher synchronisiert, welcher seinen einzigartigen Akzent imitieren konnte (Lee, 2019). In der Ausstellung «Dalí Lives» erscheint der Künstler den Besuchern, wenn sie die Türklingel an dem Kiosk drücken, in welchem er lebte, worauf der Spanier Geschichten aus seinem Leben erzählt (Lee, 2019; Mahmud & Sharmin, 2020, S. 14). Diese Art der Kunst hilft den Besuchern, den Künstler in den Kontext des modernen Lebens zu bringen und Dalí als Mensch näher kennenzulernen (Lee, 2019).

Mittlerweile können auch künstlerische Werke, wie diejenigen von Monet und Van Gogh, nachgeahmt werden (Whittaker et al., 2020, S. 97). GANs können dabei zur stilistischen Übertragung von Bildmaterial verwendet werden, indem diese auf Landschaftsfotografien trainiert werden und Sammlungen von Kunststilen nachahmen (Whittaker et al., 2020, S. 97). Dabei können die Originalfotografien im Stil des Künstlers generiert werden

und bieten somit neue Möglichkeiten, künstlerische Stile zu replizieren und neue Kunstwerke zu erschaffen (Whittaker et al., 2020, S. 97). Öffentliche Websites wie Artbreeder ermöglichen jedem Benutzer, zum GAN-Künstler zu werden und Bilder zu erstellen, um völlig neue durch GAN-Technologie konstruierte Kunstwerke zu generieren (Artbreeder, 2021).

### **5.1.3 Multimediaindustrie**

Die Film- und Videospieldustrie gehören zu den Branchen, in welchen die neue Technologie am meisten zum Einsatz kommen kann (Gardiner, 2019, S. 19). Die Multimediaindustrie, die digitale Videofiguren verwendet, erreicht hochwertige visuelle Effekte hauptsächlich durch die erneute Synthetisierung von Audio und Video (Gardiner, 2019, S. 19). Durch Einsatz von Deepfake-Technologie kann viel Geld und Zeit gespart werden (Mahmud & Sharmin, 2020, S. 14). Beim Netflix-Film «The Irishman» wurde Computer Generated Imagery (CGI) benutzt, damit Schauspieler wie Robert De Niro, Al Pacino und Joe Pesci in einigen Szenen um Jahrzehnte jünger erscheinen (Whittaker et al., 2020, S. 96). Der Einsatz von CGI liess das Budget des Films um bis zu 175 Millionen US-Dollar ansteigen (Whittaker et al., 2020, S. 96). Mit einer kostenlosen Software namens DeepFaceLab stellte ein YouTuber in nur sieben Tagen die Alterungseffekte von Netflix nach und veröffentlichte ein Video, in welchem er die originalen CGI-Szenen aus dem Film mit der Deepfake-Version der Schauspieler vergleicht (Whittaker et al., 2020, S. 96). Die Nachbildungen des YouTubers waren sehr überzeugend und wurden sogar höher gelobt als die kostspieligen CGI-Effekte (Whittaker et al., 2020, S. 96). Gleich wie bei der Kunst werden in Filmen verstorbene Personen digital wiederbelebt. Der verblüffende Auftritt des längst verstorbenen Schauspielers Peter Cushing im Film «Rogue One» im Jahre 2016 wurde durch eine geschickte Kombination aus Live-Schauspiel und technischer Zauberei ermöglicht (Chesney & Citron, 2018, S. 1770). Der Star-Wars-Beitrag zu diesem Thema setzte sich im Film «Die letzten Jedi» fort, als der vorzeitige Tod von Carrie Fisher die Filmemacher dazu veranlasste, zusätzliche Dialoge mit Schnipseln aus echten Aufnahmen zu fälschen (Chesney & Citron, 2018, S. 1770). Obwohl diese Anwendungen ethische Diskussionen auslösen, können Deepfakes Filmemachern eine kostengünstige Alternative zu CGI bieten, indem das umfangreiche Audio- und Bildmaterial verstorbener Schauspieler genutzt wird (Whittaker et al., 2020, S. 96).

Auch im Bereich der Videospiegelindustrie ist der Einsatz von Deepfake bemerkenswert (Albahar & Almalki, 2019, S. 3246). Die Technologie kann benutzt werden, um Sprachaufnahmen für In-Game-Dialoge zu generieren (Ambalina, 2020; Buo, 2020, S. 2). Spielentwickler können somit den Dialogtext eingeben, eine zur Szene passende Emotion auswählen und einen Audioclip mit dem gesprochenen Teil des Dialogs erzeugen (Ambalina, 2020; Buo, 2020, S. 2). Synthetisch generiertes Audio wird somit den Zeit- und Kostenaufwand für herkömmliche Voice-over-Aufnahmen reduzieren (Ambalina, 2020). Wenn die Technologie in den nächsten Jahren einen Punkt erreicht, wo die Emotionen präzise nachgeahmt werden können, wird dies eine grosse Veränderung in der Videospiegelindustrie im Bereich des Voice Acting herbeiführen (Ambalina, 2020). Ausserdem ermöglichen Deepfakes die Nachahmung von Mimik und Gesichtszügen von echten Personen (Schick, 2020, S. 61). Dies wird die visuelle Darstellung in Spielen wie FIFA so verändern, dass die Spielcharaktere wie die echten Sportler aussehen (Schick, 2020, S. 61).

Bei den aufgeführten Beispielen der drei Bereiche Bildung, Kunst und Multimediaindustrie handelt es sich um keine abgeschlossene Liste von positiven Anwendungen der Deepfake-Technologie. Deepfakes haben zusätzlich in den folgenden Branchen positives Potential: in der Beseitigung von Sprachbarrieren, im Gesundheitswesen und in verschiedenen Geschäftsfeldern wie Mode und E-Commerce (Albahar & Almalki, 2019, S. 3246; Westerlund, 2019, S. 41; Whittaker et al., 2020, S. 96). Ausserdem ermöglichen Apps wie Zao, FaceApp oder Reface ihren Nutzern grenzenlose personalisierte Mediengestaltung und Unterhaltung, wodurch die eigenen Kreationen mit der ganzen Welt geteilt werden können. Doch trotz der vielen positiven Anwendungen geniessen Deepfakes in der Öffentlichkeit weitgehend einen eher schlechten Ruf.

## **5.2 Bedrohungen durch Deepfakes**

Wie viele bisherige Technologien werden auch Deepfakes eingesetzt, um ein breites Spektrum an schwerwiegenden Bedrohungen zu schaffen (Chesney & Citron, 2018, S. 1771). In Anbetracht der ungeheuren Macht, in Kombination mit Zugänglich- und Verfügbarkeit, werden Deepfakes zunehmend zu einer Bedrohung auf globaler Ebene (Muna, 2020, S. 6; Whittaker et al., 2020, S. 94). Deepfakes sind mittlerweile immer schwieriger von authentischen Videos zu unterscheiden und die Barrieren für die Erstellung von Deepfakes werden durch mobile Apps und Open-Source-Software immer geringer (Muna, 2020, S. 6; Whittaker et al., 2020, S. 94). Dadurch werden kontinuierlich mehr

Deepfake-Bilder, -Audios und -Videos erstellt und geteilt (Kietzmann et al., 2020, S. 3). Somit haben Deepfakes einen grossen Einfluss auf die aktuelle soziale und virtuelle Welt und stellen eine Bedrohung für Individuen, Politik, Gerichte, Unternehmen und die allgemeine Gesellschaft dar (Muna, 2020, S. 6). Folglich ist es wichtig, die Gesellschaft über die böswillige Nutzung von Deepfakes aufzuklären. Deshalb werden in den folgenden Abschnitten die grössten Gefahren mit aktuellen Beispielen beschrieben.

### **5.2.1 KI-Pornos**

Am 11. Dezember 2017 berichtet die Autorin Samantha Cole mit dem Titel «AI-Assisted Fake Porn is here and we're all fucked» das allererste Mal über Deepfakes (Cole, 2017). Getreu den Regeln des Internets beginnt die Deepfake-Geschichte mit Pornos, wobei bereits vor Einsatz der KI das Verbreiten von gefälschten Promi-Pornos eine der Lieblingsbeschäftigungen des Internets war (Cole, 2017; Schick, 2020, S. 43). Die Deepfake-Technologie kann somit die wildesten und bisher unzugänglichsten Fantasien befriedigen, wie zum Beispiel einen Porno mit einer berühmten Schauspielerin wie Scarlett Johansson oder einer populären Persönlichkeit wie Michelle Obama (Cole, 2017; Schick, 2020, S. 251). Dabei zeigt sich, dass Deepfake-Pornos ein unbestreitbares geschlechtsspezifisches Phänomen sind und bisher kein einziges Fake-Porno-Video mit einer männlichen Person aufgetaucht ist (Ajder et al., 2019, S. 8; Schick, 2020, S. 211). Dabei besteht die Gefahr von Deepfake-Pornos für alle Frauen und diese sind unabhängig von der Erscheinungsform eine Blossstellung sowie Demütigung für die Opfer mit verheerender Wirkung (Chesney & Citron, 2018, S. 1773; Schick, 2020, S. 213). Deep-Fake-Sexvideos zwingen Personen zu virtuellem Sex, reduzieren die Opfer zu Sexobjekten und können Vergewaltigungsdrohungen in eine erschreckende virtuelle Realität verwandeln (Chesney & Citron, 2018, S. 1773; Cole, 2017). Die Videos vermitteln die Botschaft, dass Opfer nach Lust und Laune sexuell missbraucht werden können (Chesney & Citron, 2018, S. 1773). Angesichts des Stigmas von Nacktbildern, insbesondere für Frauen und Mädchen, können diese Personen unter anderem kollaterale Konsequenzen im Privatleben und auf dem Arbeitsmarkt erleiden (Chesney & Citron, 2018, S. 1773). Nach der Recherche von Ajder et al. (2019) waren im Jahre 2019 somit rund 96 Prozent aller Deepfakes im Internet Fake-Pornos.

Seit der Entdeckung der ersten Deepfake-Pornos ist die Produktion von KI-gestützten gefälschten Sexfilmen explodiert (Cole, 2018). Immer mehr Menschen erstellen gefälschte Promi-Pornos mithilfe von Machine Learning und die Ergebnisse wurden immer

überzeugender (Cole, 2018). In den Online-Foren wurden bald neue Tools wie FakeApp entwickelt und Anleitungen entworfen, um Neulinge durch den Prozess zu führen (Cole, 2018). Als die Videos anfangen, auf anderen Plattformen zu erscheinen, verbannten Firmen wie Reddit oder PornHub die gefälschten Videos, worauf im Februar 2018 die erste offizielle Deepfake-Pornoseite registriert wurde (Schick, 2020, S. 52; Westerlund, 2019, S. 44). Mittlerweile wurden die Videos auf den Top vier Deepfake-Pornoseiten über 134 Millionen Mal angeschaut (Schick, 2020, S. 52). Das Problem dieser Seiten liegt in der Einfachheit des Zugangs. Alle Seiten können durch eine einfache Google-Suche gefunden werden und sind frei zugänglich. Die Webseiten bezeichnen sich dabei als Communities von Deepfake-Pornos für Unterhaltungszwecke. Durch das Registrieren erhält man Zugang zum regen Austausch der Community, zu diversen Anleitungen zur Erstellung von Deepfakepornos sowie zum Marktplatz für personalisierte Wünsche. Ein Beispiel, wie eine Anfrage für ein personalisiertes Video aussehen kann, findet man in Abbildung 14. Dabei können registrierte User ihre Anfrage für ein personalisiertes Video einfach im Forum platzieren. Als Gegenangebot wird oft eine Bezahlung in Form von Bitcoins angeboten. Dadurch ist der Zugang zu Deepfake-Pornos sehr einfach und die Filme können mit wenigen Klicks konsumiert werden. Trotz der gefährlichen Arbeiten auf diesen Portalen müssen auch dort Regeln eingehalten und respektiert werden. Auf der grössten Deepfake-Community-Seite dürfen zum Beispiel keine Deepfake-Pornos mit nicht berühmten Personen oder von Kindern erstellt werden. Ein weiteres Gebot ist der gute Umgang innerhalb der Community. Wenn sich ein User nicht an die vorgeschriebenen sieben Regeln hält, wird er ohne Warnung von der Seite verbannt. Trotz des Verbots von Deepfake-Pornos mit nicht berühmten Personen, besteht die Gefahr für alle Frauen weiterhin. Damit diese Arbeit keine Förderung von diesen Seiten ist, werden die Quellen dazu explizit nicht genannt.

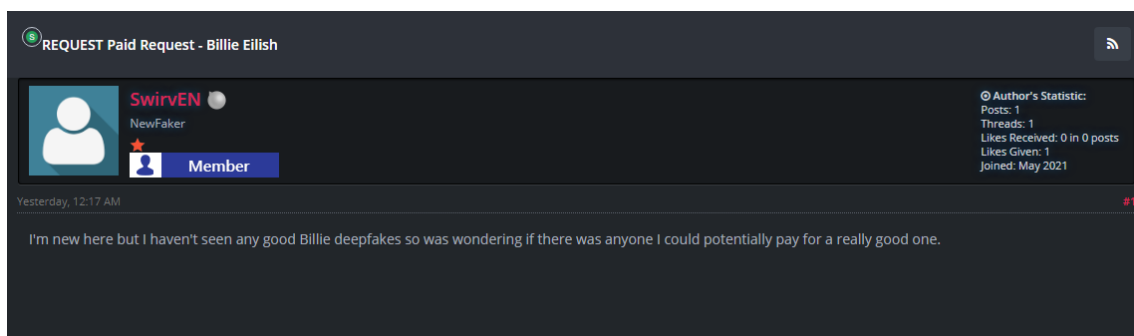


Abbildung 14: Anfrage in einem Deepfake-Forum.

Die Firma Sensity AI veröffentlichte im Oktober 2020 einen Report über die Untersuchung eines neuen Deepfake-Ökosystems auf der Messanging-Plattform Telegram. Im Mittelpunkt dieses Ökosystems steht ein KI-gesteuerter Bot, der es den Nutzern ermöglicht bekleidete Bilder von Frauen fotorealistisch in Nacktbilder umzuwandeln (Ajder et al., 2020, S. 2). Im Vergleich zu ähnlichen Tools erhöht der Bot die Zugänglichkeit dramatisch, da er eine kostenlose und einfache Benutzeroberfläche bietet, die sowohl auf Smartphones als auch auf Computern funktioniert (Ajder et al., 2020, S. 2). Wie einfach es ist, ein solches Nacktbild von einer gewünschten Person zu erhalten, wird in Abbildung 15 dargestellt. Um ein Nacktbild einer definierten Person zu erhalten, laden Benutzer einfach ein Foto einer Zielperson auf den Bot hoch und erhalten nach einem kurzen Generierungsprozess das bearbeitete Bild (Ajder et al., 2020, S. 5). Dabei wurde die verwendete GAN-Technologie auf einen Satz von Bildern von bekleideten und nackten Frauen trainiert, wodurch die Software unbegrenzt verwendet werden kann, um Fotos von zuvor ungesehenen Zielen zu entkleiden (Ajder et al., 2020, S. 5). Die Untersuchung von Ajder et al. (2020) ergab dabei, dass bis Juli 2020 Nacktbilder von ungefähr 104'852 Frauen produziert und in Umlauf gebracht wurden. Weltweit verzeichnete das Ökosystem rund 101'080 Mitglieder, wobei 70 Prozent aus Russland und den Ländern der ehemaligen UdSSR stammen.



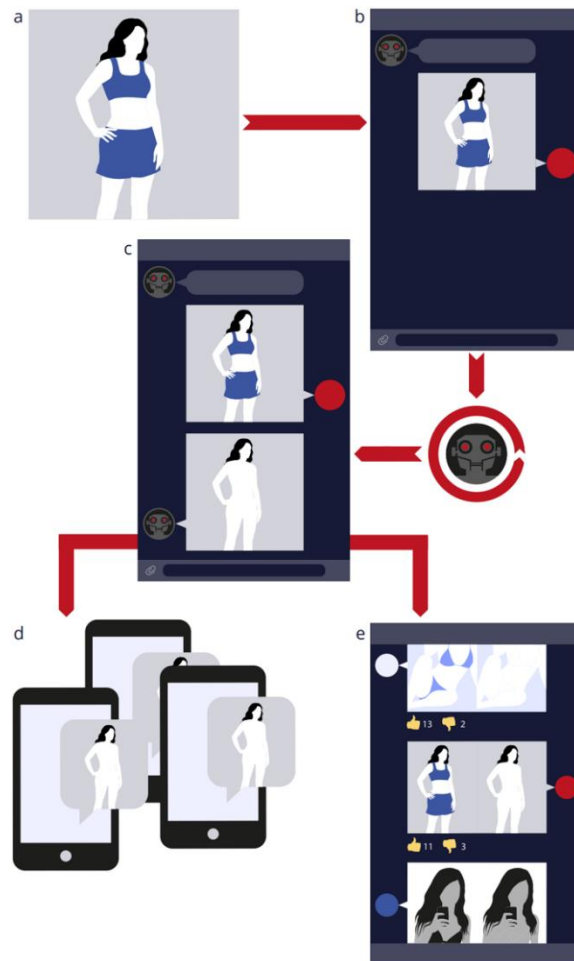


Abbildung 15: Generierung gefälschter Nacktbilder (Ajder et al., 2020, S. 4).

Wie der erste Deepfake-Porno können auch die heutigen Videos immer noch als Fake identifiziert werden, jedoch steigt die Qualität stetig an (Schick, 2020, S. 213). Wenn es darum geht, Einzelpersonen anzugreifen, ist der erste und böartigste Einsatz von Deepfakes bereits dabei. Jeder nicht-einvernehmliche Fake-Porno, ob guter oder schlechter Qualität, ist für die Opfer erschreckend, peinlich und erniedrigend. Obwohl die gefälschten Sexvideos in erster Linie auf Prominente abzielen, können Privatpersonen genauso leicht ihr Abbild in einem dieser Videos mit ausgetauschtem Gesicht wiederfinden (Delfino, 2020, S. 895). Betroffene Opfer von Deepfake-Pornos stehen vor einzigartigen Herausforderungen, wenn sie ein Mittel gegen diese böswillige Nutzung der Technologie suchen (Delfino, 2020, S. 896). Die oben genannten Beispiele sind nur einige beunruhigende, bekannte Fälle. Andere Formen von gefälschten sexuellen Inhalten werden folgen (Chesney & Citron, 2018, S. 1774).

### **5.2.2 Verfälschung der Realität**

Deepfake-Pornos sind der Beweis dafür, dass die KI-gestützte Technologie verwendet werden kann, um den Ruf von berühmten oder nicht bekannten Personen zu schädigen (Masood et al., 2021, S. 2). Nebst dieser böswilligen Nutzung gegen bestimmte Persönlichkeiten stellen diese Videos jedoch auch im Grunde eine falsche Tatsache und somit eine Fehlinformation dar (Masood et al., 2021, S. 2). Bei jedem einzelnen Deepfake handelt es sich grundsätzlich um eine Verfälschung der Wahrheit (Schick, 2020, S. 14). Dadurch können Deepfakes einen grossen gesellschaftlichen Schaden anrichten und zu einem Krieg um die Realität führen (Brown, 2020, S. 10). Einige Betrachter von Deepfakes könnten als Reaktion auf die falschen Informationen Massnahmen ergreifen, bevor die Fälschung des Inhalts bekannt geworden ist (Brown, 2020, S. 10). Wiederum könnten andere die Beweise ablehnen, die zeigen, dass das Filmmaterial gefälscht ist, insbesondere wenn der Inhalt die Vorurteile oder bereits bestehenden Annahmen des Betrachters bestätigt (Brown, 2020, S. 10). Somit wird ein Widerruf einige Leute nicht davon abhalten, das gefälschte Video zu glauben und zu verbreiten (Brown, 2020, S. 10). In zunehmendem Masse nutzen bösartige Akteure von Nationalstaaten bis hin zu Einzelpersonen diese neuen Umstände, um Des- und Fehlinformationen für die eigenen schädlichen Zwecke zu verbreiten und die Meinung der Gesellschaft zu beeinflussen (Schick, 2020, S. 13). Die Weiterentwicklung von Deepfakes und der einfache Zugang zur Technologie stellen eine Gefahr für das gesamte Informationsökosystem dar (Schick, 2020, S. 11). Aufgrund dieser Entwicklung und zunehmender Bedrohung werden auf den folgenden Seiten aktuelle Beispiele von Des- und Fehlinformationen in der Politik, bei Unternehmen und Einzelpersonen beschrieben.

#### **5.2.2.1 Politische Manipulationen**

Deepfakes können zur Ausnutzung für politische Vorteile wie zum Beispiel bei Wahlsystemen oder Rufschädigung einer anderen politischen Person führen (Muna, 2020, S. 8). Durch Deepfakes können gefälschte Videos in sozialen Medien oder Nachrichtenkanälen verbreitet werden, die eine politische Figur in einem schlechten Licht handelnd zeigen (Muna, 2020, S. 8). Dadurch werden die öffentliche Wahrnehmung der besagten politischen Person und die politische Kampagne verändert, was zwangsläufig die Wahl beeinflusst und die öffentliche Wertschätzung der Person verringert (Muna, 2020, S. 9). Im Mai 2019 postete ein Trump-Supporter auf seiner persönlichen Facebook-Seite ein Videoclip von Nancy Pelosi, der Sprecherin des US-Repräsentantenhauses (Poulsen, 2019).

Das Video zeigt die Politikerin während einer Pressekonferenz, bei welcher sie sich über ein gescheitertes Infrastruktur-Meeting mit dem Präsidenten aufregt, jedoch lässt das manipulierte Video Nancy Pelosi betrunken erscheinen (Poulsen, 2019). Bevor man die Manipulation der Geschwindigkeit nachweisen konnte, wurde das Video mehr als 2.5 Millionen Mal in den Social-Media-Kanälen angeschaut (Greengard, 2019, S. 18; Poulsen, 2019). Der damalige US-Präsident Donald Trump postete das Video auf seinem Twitter-Kanal, um das Image der Politikerin zu schwächen (Greengard, 2019, S. 18; Poulsen, 2019). Dieser einfache Deepfake zeigt die grossen Auswirkungen auf politische Figuren, indem einige getäuscht werden, während andere wiederum das Video zum eigenen Vorteil benutzten.

In der Politik sind Wahlen auf eine besondere Weise anfällig für Fälschungen. In einer Demokratie ist der Diskurs am funktionalsten, wenn Debatten auf einem Fundament gemeinsamer Fakten und Wahrheiten aufgebaut sind, welche durch empirische Beweise gestützt werden (Chesney & Citron, 2018, S. 1778). Dabei können Deepfakes als ein weiteres nützliches Werkzeug verwendet werden, um durch staatlich geförderte Desinformationskampagnen Wahlen zu stören und Unruhe zu stiften (Whittaker et al., 2020, S. 95). Nach den Forschungen von Schick (2020, S. 80) führt Russland bei den Präsidentschaftswahlen der Vereinigten Staaten von Amerika im Jahr 2016 eine lang geplante Desinformationskampagne durch. Dabei begann Russland bereits drei Jahre vor der Wahl, den öffentlichen Diskurs in den USA zu infiltrieren, indem sie sich in den sozialen Medien als authentische Amerikaner ausgaben. Anschliessend verbreiteten diese Fake-Profile so viel Uneinigkeit, Polarisierung, Spaltung und Desinformation wie möglich. Die Russen nutzten die Schwachstellen der USA aus, indem sie die Gesellschaft von innen heraus spalteten und Trump und Clinton-Supporter gegeneinander aufhetzten. Obwohl die etlichen gefälschten Profile und Communities bewiesen werden konnten, fehlt bis heute der Nachweis über die Involvierung und Auftragserteilung durch die russische Regierung. Heutzutage gibt es viele Länder, welche von der Regierung gesponserte Social-Media-Accounts, Websites und Anwendungen betreiben und so zu politischer Propaganda auf der ganzen Welt beitragen (Masood et al., 2021, S. 3). Insbesondere die Regierungen Chinas, Israels, der Türkei, Russlands, Grossbritannien, der Ukraine und Nordkorea sind daran beteiligt, digitale Kampagnen zu betreiben, um Gegner zu diffamieren, Desinformationen zu verbreiten und gefälschte Texte zu veröffentlichen (Masood et al., 2021, S. 3).

Durch die Weiterentwicklung von Deepfakes können diese in Zukunft eine noch grössere politische Gefahr werden, indem Videos von Präsidenten mit militärischen Aussagen gegen ihre Gegner auftauchen könnten. Bevor die Manipulation der Videos entdeckt wird, können politische Führer oder Armeegeneräle sich gezwungen sehen, schnell zu reagieren. In einem solchen Schreckensszenario kommt es zu einem Dominoeffekt, der möglicherweise zu grossen Kriegen führt, die alle auf ein Deepfake-Video zurückzuführen sind. Selbst wenn ein solches Szenario sehr dramatisch klingt, ist die Idee, dass Deepfakes in Zukunft Kriege auslösen können, schwer zu bestreiten.

#### **5.2.2.2 Gefahr für Unternehmen**

Unternehmen und Organisationen im Bereich der privaten Wirtschaft können mit wenig Aufwand durch die Deepfake-Technologie beeinflusst werden (Muna, 2020, S. 9). Unternehmen mit einer bedeutenden Stimme oder Rolle für die Gesellschaft auf lokaler oder nationaler Ebene werden durch die Technologie vermehrt Angriffen auf ihren Ruf ausgesetzt sein (Chesney & Citron, 2018, S. 1779). In der Anfangszeit werden die meisten Firmen Opfer von Betrugereien werden (Kietzmann et al., 2020, S. 143). Im Moment sind Deepfakes im Bereich des Betrugs gegen Firmen noch nicht die Norm, sie können jedoch in Zukunft zu einem beliebten Werkzeug für viele Hacker werden, welche die Ausrüstung und Geduld haben, die Technologie für sich zu nutzen (Sjouwerman, 2020). Indem Schlüsselfiguren in einer Organisation nachgeahmt werden und CEOs in belastende Situationen hinein manipuliert werden, können sich Deepfakes unweigerlich auf Verhandlungen zwischen Unternehmen oder Partnerschaften auswirken (Muna, 2020, S. 9; Sjouwerman, 2020). Ausserdem können sich Kriminelle durch die neue Technologie leicht als Vorgesetzte eines Unternehmens ausgeben, um Mitarbeiter zu einer unbedachten Handlung aufzufordern (Sjouwerman, 2020).

Im März 2019 geschah in Europa die erste bekannte Attacke auf ein Unternehmen durch ein von KI generiertes Sprach-Deepfakes (Stupp, 2019). Die Kriminellen setzten dabei eine Software ein, um sich als die Stimme eines Geschäftsführers auszugeben und erbeuteten dabei eine Summe von 220'000 Euro (Sjouwerman, 2020; Stupp, 2019). Gemäss dem Bericht von Stupp (2019) glaubte der Geschäftsführer eines britischen Energieunternehmens, mit seinem Chef der deutschen Muttergesellschaft zu telefonieren. Dieser bat das Opfer, dringend Geld an einen ungarischen Lieferanten zu überweisen. Der britische

CEO erkannte dabei die Sprachmelodie und den leichten deutschen Akzent seines Chefs und überwies den geforderten Betrag auf ein ungarisches Bankkonto. Jedoch handelte es sich bei dem Anruf um eine Täuschung, bei welchem die Betrüger eine KI-basierte Software verwendeten, um einen Sprach-Deepfake des deutschen Geschäftsführers am Telefon zu imitieren. Solche Betrügereien sind eine neue Herausforderung für Unternehmen, da die bisherigen Cybersicherheits-Tools gefälschte Stimmen nicht erkennen.

Deepfakes können auch in den sozialen-Netzwerken eine Gefahr für Firmen und Organisationen darstellen, wie der Fall von Maisy Kinsley zeigte. Im März 2019 drohte der Kampf zwischen der Firma Tesla und Aktien-Leerverkäufern zu eskalieren, wobei es um Milliarden ging. In dieser Zeit vernetzte sich die angeblich leitende Bloomberg-Journalistin mit 195 Investoren von Tesla-Aktien auf LinkedIn (Schick, 2020, S. 207). Ausserdem begann Maisy Kinsley prominenten Tesla-Leerverkäufern auf Twitter zu folgen (Schick, 2020, S. 208). Sie begann schliesslich, mehrere Personen anzuschreiben und bat diese um persönliche und finanzielle Informationen (Fleishman, 2019). Einer der angeschriebenen Aktien-Leerverkäufer sah sich das Profil von Kinsley genauer an und wurde misstrauisch (Fleishman, 2019). Obwohl es eine professionelle Webseite und ein LinkedIn-Profil gab, erschienen auf der Bloomberg-Webseite keine Berichte unter ihrem Namen, was für eine angebliche leitende Journalistin ungewöhnlich war (Fleishman, 2019). Schliesslich stellte sich heraus, dass Maisy Kinsley gar keine Journalistin war, sondern ein gefälschtes Profil, wobei das Profilbild ein mittels GAN-Technologie synthetisiertes Foto war (Fleishman, 2019; Schick, 2020, S. 208). Bloomberg bestätigte später, dass sie keine Mitarbeiterin namens Maisy Kinsley haben (Fleishman, 2019). Wie der Fall von Maisy zeigt, können Deepfakes auch dazu verwendet werden, sensible Investitionsinformationen zu erlangen (Schick, 2020, S. 209). Obwohl begrenzter Schaden entstanden ist und nur ein Deepfake-Foto benutzt wurde, könnten in Zukunft solche Fake-Profile mit Deepfake-Videos ergänzt werden, um das Vertrauen der Zielpersonen zu gewinnen und an die benötigten Informationen zu kommen.

Wenn sich die Deepfake-Technologie in dem gleichen Tempo weiterentwickelt wie bisher, wird es für Kriminelle bald möglich sein, komplett fiktive Geschichten für Fake-Profile zu erstellen. Betrug und Erpressung sind seit Langem Aktivitäten, nach denen Kriminelle streben. Deepfakes werden es Personen mit böswilligen Absichten nur noch

einfacher machen, Firmen und Organisationen zu täuschen, wobei die Kosten und Verluste für die Unternehmen in diesem Zusammenhang gewaltige finanzielle Ausmasse annehmen werden.

### **5.2.2.3 Täuschung und Diffamierung von Einzelpersonen**

Nebst Staaten und Unternehmungen können auch Einzelpersonen durch Deepfakes getäuscht oder öffentlich geschädigt werden. Eine Bande von Betrügern schaffte es 2016, insgesamt 50 Millionen Euro von einigen der klügsten und mächtigsten Persönlichkeiten Frankreichs zu erbeuten, indem sie sich als damaligen französischen Verteidigungsminister Jean-Yves Le Drian ausgaben (Breedon, 2020; Schick, 2020, S. 189). Dabei verliessen sie sich auf die Macht der audiovisuellen Kommunikation und kontaktierten etliche wohlhabende Personen über Telefon- und Videoanrufe (Breedon, 2020). Dabei baten sie ihre Opfer, Geld auf polnische und chinesische Konten zu überweisen, um damit streng geheime Missionen der französischen Regierung zu finanzieren, bei welcher es um die nationale Sicherheit geht (Breedon, 2020). In einigen Fällen meldeten sich die Betrüger sogar mit einem kurzen unscharfen Videoanruf, in dem ein Mann dank einer Silikonmaske wie Le Drian aussah (Breedon, 2020). Bei dem dreisten Betrug wurden damals einfache Mittel eingesetzt, welche mit der heutigen Deepfake-Technologie wenig zu tun hatten. Trotzdem war die schlechte Imitation der Betrüger gut genug, um einige der reichsten Männer der Welt zu überzeugen, sich von Millionen zu trennen.

Erniedrigung von Einzelpersonen durch Deepfakes geschieht in den meisten Fällen, wie in Kapitel 5.2.1 erwähnt, durch nicht-einvernehmliche pornografische Inhalte. Nebst Politiker-, Sänger- und Schauspielerinnen können oft auch Personen, die nicht im Rampenlicht der Öffentlichkeit stehen, zur Zielscheibe werden, wie der Fall der 18-jährigen Frau, die aus Neugierde eine umgekehrte Bildsuche durchführte und dabei hunderte Bilder mit ihrem Gesicht in pornografischen Szenen fand (Melville, 2019; Whittaker et al., 2020, S. 95). Gemäss dem Bericht von Melville (2019) lud die 18-jährige Noelle Martin ein Bild von sich selbst hoch, um zu schauen, wo im Internet noch weitere Bilder von ihr vorhanden sind. Anstatt wie erwartet weitere Bilder auf den Social-Media-Kanälen ihrer Freunde zu finden, überschwemmten hunderte freizügige Fotos, bei denen ihr Gesicht auf die Körper von Pornodarstellerinnen montiert wurde, ihren Bildschirm. Nachdem die Australierin begann, sich öffentlich darüber zu äussern, wurde sie von einer breiten Internetcommunity angefeindet. Weitere gefälschte Bilder erschienen, bis schliesslich ein Deepfake-

Pornovideo mit ihrem Gesicht auf verschiedenen Webseiten hochgeladen wurde. Die Täter fühlten sich durch den öffentlichen Diskurs, welcher durch Noelle Martin ausgelöst wurde, angegriffen und wollten sie mit weiteren Deepfakes zum Schweigen bringen. Heute wird Noelle Martin noch immer im Internet belästigt.

Bis heute bleiben Frauen die Hauptopfer von böswilligen Deepfakes. Obwohl viele Online-Plattformen Deepfakes verbannt haben, werden die erniedrigenden Inhalte weiterhin im Internet verbreitet und angeschaut. Zusätzlich können Deepfakes einen schädlichen Einfluss auf die Geopolitik und die Beziehungen zwischen Ländern haben. Gefälschte Videos und synthetische Inhalte auf den sozialen Netzwerken beeinflussen dabei ganze Gesellschaften und spalten die Bevölkerung. Firmen und Organisationen müssen sich zunehmend vor Betrügen durch kriminelle Organisationen schützen, welche die neuen Technologien für feindliche Absichten nutzen. Die in diesem Kapitel beschriebenen Vorfälle und Gefahren sind eine kleine Aufzählung von bekannten Ereignissen und Bedrohungen, welche durch Deepfakes entstanden sind. Durch die Weiterentwicklung in den nächsten Jahren kann es möglich werden, dass die Technologie öfters für feindliche Absichten eingesetzt wird. Damit der Kampf um die Wahrheit und die Realität nicht verloren wird, müssen technologische wie auch rechtliche Mittel entwickelt werden, um böswillige Deepfakes effizient bekämpfen zu können.

## **6 Lösungen zur Bekämpfung von feindlichen Deepfakes**

Die durch KI neu entstandene Deepfake-Technologie ist für sich genommen selbstverständlich weder gut noch schlecht. Wie in Kapitel vier beschrieben, kann die Technik für gute sowie auch für böswillige Zwecke eingesetzt werden. Jedoch überwiegt aktuell der negative Schaden den positiven Nutzen, da die Deepfake-Technik weitreichend sozialen Schaden anrichten kann (Farid & Schindler, 2020, S. 29). Bei Betrachtung der Tatsache, dass die Technik eine hohe Zahl an Nutzern erreichen kann sowie Hard- und Softwareanforderungen keine grossen Probleme mehr darstellen, sollte eine Reihe von Massnahmen, welche den sozialen Schaden minimieren können, entwickelt werden (Farid & Schindler, 2020, S. 29). Deshalb werden in diesem Kapitel die technischen und rechtlichen Ansätze erläutert und auf deren Probleme hingewiesen.

## 6.1 Technische Lösungen

Seitdem Deepfakes in der Öffentlichkeit aufgetaucht sind, liefern sich die Angreifer und Verteidiger ein Wettrüsten. Mit der rasanten Entwicklung der Deepfake-bezogenen Studien haben beide Seiten eine Art Schlachtfeld gebildet, welches die Verbesserungen der jeweils anderen Seite vorantreibt und neue Richtungen anregt (Juefei-Xu et al., 2021, S. 1; Mirsky & Lee, 2021, S. 32). Eine effiziente und allgemein wirksame Methode zur Erkennung von Fälschungen zu entwickeln ist eine gewaltige Herausforderung (Chesney & Citron, 2018, S. 1787). Die Methode muss mit den Innovationen in der Deepfake-Technologie mithalten können, um wirksam zu bleiben, und zugleich in der Lage sein, nützliche Anwendungen uneingeschränkt fortbestehen zu lassen (Chesney & Citron, 2018, S. 1787). Von der Wissenschaft werden aktuelle Studien über das Wettrüsten und den gegenwärtigen Stand von technischen Lösungsansätzen gegen böswillige Deepfakes vernachlässigt (Juefei-Xu et al., 2021, S. 1). Da die Zahl der Publikationen rapide angestiegen ist in den letzten Jahren, ist ein tiefgreifendes Verständnis der Tendenz und der zukünftigen Arbeiten beeinträchtigt (Juefei-Xu et al., 2021, S. 1). Die Arbeit von Juefei-Xu et al. (2021, S. 21) liefert als erste einen umfassenden Überblick der Forschungsarbeiten zum Thema Deepfake-Erkennung und Deepfake-Erzeugung, wobei über 191 Arbeiten analysiert wurden. Dabei zeigt sie das Wettrüsten zwischen den beiden Parteien mit detaillierten Interaktionen zwischen den Deepfakeerzeugungs- und Deepfakeerkennungsmethoden. Dadurch wurde eine neue Perspektive auf die aktuelle Lage der Deepfake-Forschung ermöglicht, welche wertvolle Informationen zu den Herausforderungen und Möglichkeiten für die Zukunft liefern kann.

Die Arbeit von Mirsky & Lee (2021, S 32) zählt die wichtigsten Verbesserungen und Errungenschaften im Bereich der Erkennungsmethoden auf. Wichtige Fortschritte in diesem Bereich waren die Identifizierung von Artefakten, die während des Erstellungsprozesses zurückgelassen wurden, wie Unstimmigkeiten in der Kopfhaltung, fehlendes Augenblinzeln, Farbvariationen in der Gesichtstextur, die Ausrichtung der Zähne, räumlich-zeitliche Merkmale sowie die Verhaltensmuster einer Person. Die Arbeiten von Mirsky & Lee (2021) sowie Juefei-Xu et al. (2021) sind beide zum Schluss gekommen, dass im Bereich der automatischen Erkennung von Deepfakes immer noch Verbesserungsbedarf vorhanden ist. Die meisten Arbeiten haben sich auf die Erkennung von Face-Swaps konzentriert, jedoch werden durch die immensen Entwicklungen andere Kategorien von Deepfake wie die Nachahmung oder Lip-Synching immer stärker. Ausserdem existieren



nur wenige Arbeiten, die sowohl Audio- als auch visuelle Deepfakes erkennen können. Die meisten Algorithmen zur Deepfake-Erkennung gehen von einem statischen Wechselspiel mit dem Angreifer aus, wodurch sie entweder auf die Erkennung eines bestimmten Artefaktes fokussiert sind oder sich nicht gut auf ungesehene Angriffe anwenden lassen. In den meisten Studien zu Erkennungsmethoden wurde ausserdem versäumt, die Effektivität der Methode bei der Bekämpfung von ungesesehenen Deepfakes zu evaluieren, was für einen Detektor in der freien Wildbahn entscheidend ist.

Da die Qualität synthetisch hergestellter Inhalte immer besser wird, werden die Menschen eines Tages nicht mehr in der Lage sein, Deepfakes mit blossen Augen zu erkennen. Aus diesem Grund hat die Wissenschaft seit der Entdeckung von Deepfakes intensiv an Erkennungsmethoden geforscht und auch einige Ergebnisse erzielt. Jedoch wurde damit auch ein sich ständig weiterentwickelndes Katz-und-Maus-Spiel zwischen Erstellungs- und Erkennungsmethoden geschaffen. Je besser die Tools gegen Deepfakes werden, desto besser werden auch die Deepfakes. Laut einigen Experten ist die Forschung noch einige Jahre davon entfernt, eine Technologie entwickelt zu haben, welche einen echten Inhalt von einer Täuschung unterscheiden kann (Chesney & Citron, 2018, S. 1788; Juefei-Xu et al., 2021, S. 26; Masood et al., 2021, S. 22). Neue Entwicklungen in den Bereichen der rechtlichen Regelungen sowie ein wachsendes Bewusstsein könnten nebst den technologischen Tools wichtige Mittel im Kampf gegen böswillige Deepfakes darstellen.

## **6.2 Rechtliche Lösungen**

Neben den Entwicklungen von technischen Abwehrlösungen gegen Deepfakes besteht in jedem Fall auch in juristischer Hinsicht Handlungsbedarf. Eine Recherche zeigte auf, dass es in der Schweiz bisher noch keine gesetzlichen Regelungen gibt, welche eine eindeutige Grenze zwischen einer zulässigen Bearbeitung von Videos und deren Verbreitung sowie unzulässigen Täuschungen zieht. Diese Problematik existiert nicht nur in der Schweiz, sondern herrscht in vielen Ländern weltweit. Mit Blick auf die wachsende Problematik des Themas ist es daher notwendig, dass sich der Gesetzgeber mit dem Thema auseinandersetzt, damit betroffene Personen besser geschützt werden können. Gemäss Chesney und Citron (2019, S. 1789) müssen jedoch beim Entwurf solcher gesetzlichen Richtlinien einige Punkte beachtet werden. Ein pauschales Verbot gegen Deepfakes würde keinen Sinn ergeben, da digitale Manipulationen an sich nicht problematisch sind. Ein Verbot

von Deepfakes würde Routinemodifikationen, welche die Verbesserung und Klarheit digitaler Inhalte vorantreibt, verhindern. In den bereits erwähnten Bereichen Bildung, Kunst, Multimediaindustrie, Gesundheitswesen und Wissenschaft würde dies die potenziellen Chancen der Technologie beeinträchtigen. Ein Gesetz zu entwerfen, welches schädliche Anwendungen verbietet und gleichzeitig nützliche ausschliesst, ist schwierig zu formulieren, aber nicht unmöglich. In einigen Ländern auf der Welt wurden bereits Gesetze gegen Deepfakes erlassen.

In den Vereinigten Staaten von Amerika wurden explizite Gesetze bezüglich Deepfakes in das Strafrecht aufgenommen (Ferraro, 2019, S. 2). Als erster Staat führte Virginia im März 2019 strafrechtliche Sanktionen für die Verbreitung von nicht einvernehmlicher Deepfake-Pornografie ein (Ferraro, 2019, S. 15; Langa, 2021, S. 777). Das Virginia-Gesetz machte die Verbreitung von nicht einvernehmlich synthetisch generierten Nacktbildern und -videos zu einem Vergehen, das mit bis zu einem Jahr Gefängnis und einer Geldstrafe von \$2'500 US-Dollar bestraft wird (Ferraro, 2019, S. 15; Langa, 2021, S. 777). Es folgten weitere Gesetze in anderen Teilen des Landes, wobei Kalifornien und Texas als einzige Bundesstaaten rechtliche Vorschriften zum Verbot von Deepfakes zur Verhinderung von Wahlbeeinflussung erlassen haben (Ferraro, 2019, S. 14; Langa, 2021, S. 775). Dabei enthielt vor allem die kalifornische Gesetzgebung eine wesentlich detaillierte Definition darüber, was Deepfakes ausmacht. Das Gesetz verbietet einer Person oder einem Komitee, innerhalb von 60 Tagen vor einer Wahl mit tatsächlicher Boshaftigkeit irreführende Audio- oder visuelle Medien über den Kandidaten zu verbreiten (Langa, 2021, S. 776). Dabei muss die Verbreitung die Absicht verfolgen, den Ruf des Kandidaten zu schädigen oder einen Wähler zu täuschen, damit dieser für oder gegen den Kandidaten stimmt (Langa, 2021, S. 776). Das Gesetz bezieht sich nicht ausdrücklich auf Deepfakes, sondern verbietet lediglich Bilder, Audio- und Videoaufnahmen, die einer rationalen Person fälschlicherweise als realistisch erscheinen würden und diese Person dazu veranlassen würden, ein grundlegend anderes Verständnis von dem Ausdrucksinhalt zu haben (Langa, 2021, S. 776). Gleichzeitig sieht das Gesetz Ausnahmen für manipulierte Medien vor, die Satire oder Parodie darstellen (Langa, 2021, S. 776).

Nebst den Vereinigten Staaten von Amerika hat der chinesische Gesetzgeber die möglicherweise tiefgreifendste Regulierung von Deepfakes vorgenommen, welche am 1. Januar 2020 in Kraft getreten ist. Das chinesische Gesetz verlangt von den Anbietern und Nutzern von Online-Videonachrichten- und Audio-Informationsdiensten eine eindeutige

Kennzeichnung aller Inhalte, die mit neuen Technologien wie künstlicher Intelligenz erstellt oder verändert wurden (Yang & Goh, 2019). Parallel dazu existiert ein Verbot der Veröffentlichung und Verbreitung von Fehl- und Desinformationen, die mit Technologien wie KI und virtueller Realität erstellt wurden (Yang & Goh, 2019). Dieses Gesetz hinterlässt im Fall China jedoch einen schlechten Nachgeschmack, da solche neuen Regelungen als Vorwand für weitere Zensur genutzt werden könnten.

Für die Gesetzgeber wird es eine Herausforderung sein, eine individuelle Verantwortlichkeit für schädliche Deepfakes zu erreichen, jedoch sind die Ersteller nicht die einzigen Parteien, welche die Verantwortung tragen könnten (Chesney & Citron, 2018, S. 1796). Angesichts der Schlüsselrolle, die Social-Media-Plattformen bei der Verbreitung von Deepfakes spielen, kann eine der effizientesten und effektivsten Methoden der Schadensbegrenzung darin bestehen, die Plattformen in die Haftung zu nehmen (Chesney & Citron, 2018, S. 1796; Kietzmann et al., 2020, S. 145). Online-Plattformen besitzen durch die Wirkung moralischer Überzeugung, der Marktdynamik und des potenziellen Drucks bereits einen Anreiz, Inhalte zu überprüfen, und können es sich in der Zeit von Deepfakes nicht mehr erlauben, einen Laissez-faire-Ansatz zu verwenden (Chesney & Citron, 2018, S. 1796; Kietzmann et al., 2020, S. 145). Der politische Druck hat in den USA dazu geführt, dass mehrere grosse Social-Media-Plattformen für 2019 und 2020 neue Strategien gegen die Verbreitung von Deepfakes angekündigt haben (Farid & Schindler, 2020, S. 36). Bei der Änderung seiner Richtlinien hat Facebook eine sehr enge Definition dargelegt, die nur die fortschrittlichsten Deepfake-Videoproduktionen abdeckt (Farid & Schindler, 2020, S. 37). TikTok hingegen erfasste eine sehr weit gefasste Definition, wodurch ein potenziell breites Spektrum böswilliger Aktivitäten und Formen von Deepfake-Material betroffen sein kann (Farid & Schindler, 2020, S. 37). YouTube versprach der Öffentlichkeit, dass es keine gefälschten Videos im Zusammenhang mit den Wahlen und der Volkszählung in den USA im Jahr 2020 tolerieren werde, welche die Öffentlichkeit in die Irre führen könnten (Farid & Schindler, 2020, S. 25). Es ist positiv, dass die Plattformen die Notwendigkeit entsprechender Massnahmen erkannt haben, jedoch werden aufgrund der Uneinheitlichkeit die Abwehrmechanismen eher wirkungslos bleiben (Farid & Schindler, 2020, S. 39).

Nebst den aufgeführten technischen und rechtlichen Ansätzen braucht es zudem ein allgemeines Verständnis über die neue Technologie in der Gesellschaft (Schick, 2020, S. 261). Obwohl es bereits Dutzende Unternehmen und Organisationen über die Aufklärung von Deepfakes gibt, fehlt weltweit immer noch ein gemeinsamer konzeptioneller Rahmen und eine Systematik, auf welche sich die Länder berufen können (Schick, 2020, S. 261). Bevor die Bedrohungen von Deepfakes durch technische und rechtliche Lösungen bekämpft werden können, muss ein klarer und konsistenter konzeptioneller Rahmen geschaffen werden (Schick, 2020, S. 261). Dadurch, dass die Technologie in einigen Bereichen auch eine grosse Chance darstellt, ist es schwierig technische und rechtliche Lösungen gegen die Bedrohung zu formulieren, ohne dabei den Nutzen von Deepfakes zu unterdrücken. Die Zeit für eine Diskussion um das Thema drängt, da die Erstellung von Deepfakes immer einfacher, die Inhalte immer überzeugender und die Zugänglichkeit zu den Tools leichter wird.

## **7 Erstellung eines eigenen Deepfakes**

In den vergangenen drei Jahren haben sich die Tools zur Erstellung von Deepfakes weiterentwickelt (Li et al., 2021, S. 1; Zhang et al., 2020, S. 69). Die deutliche Reduzierung der Rechenkosten sowie die leichte Verfügbarkeit von Open-Source-Tools ermöglichen mittlerweile ungeschulten Personen die Erstellung von eigenen Deepfakes (Li et al., 2021, S. 1; Whittaker et al., 2020, S. 94). In diesem Kapitel wird eine der bekanntesten und weitverbreitetsten Deepfake-Software benutzt, um einen eigenen synthetischen Inhalt herzustellen. Dabei wird das gesamte Kapitel in vier Sektionen unterteilt, in welchen die Arbeitsschritte zur Generierung eines eigenen Deepfakes beschrieben sowie die Ergebnisse reflektiert werden.

### **7.1 Vorbereitung**

Bevor mit der Erstellung des Deepfakes begonnen werden konnte, mussten einige Vorarbeiten abgeschlossen sein. Ohne diese Grundlagen hätte der Prozess für die Fertigung eines eigenen Deepfakes nicht starten können. Deshalb werden in diesem Kapitel die Vorbereitungen und die damit verbundenen Entscheidungen erläutert.

### **7.1.1 Voraussetzungen**

Bedingung für die Erstellung eines Deepfakes ist eine schnelle Grafikkarte (GPU) oder ein schneller Prozessor (CPU). Falls die nötige Hardware nicht vorhanden ist, kann die Software auch über Google Colab benutzt werden. Diese Anwendung ermöglicht es die Software zur Erstellung von Deepfakes über die Cloud arbeiten zu lassen, wobei Google bis zu 12 Stunden KI-Training am Stück erlaubt. Die benötigte Hardware für die Generierung des eigenen Deepfakes wurde durch die Zürcher Hochschule für Angewandte Wissenschaften (ZHAW) zur Verfügung gestellt.

### **7.1.2 Auswahl der Open-Source-Software**

Als erstes muss die Auswahl für die geeignete Open-Source-Software getroffen werden. Der Literatur-Review sowie Deepfake-Foren haben gezeigt, dass vor allem zwei Anwendungen in der Erstellung von Deepfake-Videos dominieren: FaceSwap und DeepFaceLab (DFL). Beide Programme sind auf GitHub erhältlich und können kostenlos heruntergeladen werden. Für die Erstellung eines eigenen Deepfakes fiel die Entscheidung schliesslich auf DFL. Die Anwendung hat sich in der Öffentlichkeit als sehr beliebt erwiesen, da viele Künstler auf YouTube DFL-basierte Videos veröffentlichen, die insgesamt über 100 Millionen Mal angeschaut wurden (Perov et al., 2020, S. 3). Ausserdem ist DFL auf dem berühmtesten und grössten Deepfake-Forum die meistbenutzte Software und es existieren einige sehr ausführliche Anleitungen zur Erstellung von Deepfakes durch DFL. Die Entwickler der Software stellen ihre Software zusätzlich in einer wissenschaftlichen Arbeit vor (Perov et al., 2020).

### **7.1.3 Installation der Software**

Die Installation der Open-Source-Software erfolgt auf der GitHub-Seite unter dem entsprechenden Repository (DeepFaceLab, 2018/2021). Dabei gibt es verschiedene Möglichkeiten die Software zu installieren. Für diese Arbeit wurde dafür der Torrent-Client «BitTorrent» verwendet (BitTorrent, 2021). Durch den Magnet Link auf der Webseite startet der Ladevorgang in BitTorrent. Bei Magnet-Links handelt es sich um einen URI-Standard für Hyperlinks, welche es erlauben auf eine oder mehrere Dateien zu verweisen, ohne darauf Rücksicht nehmen zu müssen, wo diese gespeichert werden oder dass der Verweis ungültig ist (Whitson, 2012). Der Dateienordner besitzt eine Grösse von 58.2 GB mit 134'488 Dateien aus diesem Grund wurde für den Speicherort eine externe Festplatte von WD Elements benutzt. Die heruntergeladene exe-Datei extrahiert und installiert das Programm an einem ausgewählten Ort.

#### **7.1.4 Videomaterial**

Damit der Deepfake in DFL erstellt werden kann, braucht es zwei Videodateien. Das Ursprungsvideo, welches anschliessend bei der Erstellung des Deepfakes manipuliert wird, ist durch die Zürcher Hochschule für Angewandte Wissenschaften zur Verfügung gestellt worden. Das Video ist original 20:51 Minuten lang und zeigt Professor in Wirtschaftsinformatik Dr. Thomas Keller in einem Vortrag über Mixed-Reality. Für die weitere Verwendung des Videos wird ein kleiner Ausschnitt verwendet, dieser Schritt wird jedoch im nächsten Kapitel genauer beschrieben. Das Video mit dem Zielgesicht muss zuerst erstellt werden. Insgesamt werden drei Videos mit einer Canon Mark II gefilmt. Die Videos zeigen den Autor mit verschiedenen Kopfposen und von allen Seiten sowie unterschiedlicher Mimik. Das Videomaterial wird ebenfalls auf der externen Festplatte in einem separaten Ordner abgespeichert. Damit im Rahmen dieser Arbeit bei der Erstellung des Deepfakes niemand fremden geschadet wird und keine Persönlichkeitsrechte verletzt werden, ist die Auswahl der Protagonisten in den Videos explizit auf den Autoren und den Betreuer gefallen.

#### **7.1.5 Software kennenlernen**

Nachdem die Anwendung installiert und auf der externen Festplatte gespeichert wird, ist der Aufbau und die einzelnen Funktionen und Programme der Software angeschaut worden. Die Ordner und verschiedenen Dateien zur Ausführung der einzelnen Schritte sind in Abbildung 16 dargestellt. Dabei besteht DFL aus mehreren bat-Dateien, die zur Durchführung der verschiedenen Aufgaben zur Erstellung eines Deepfakes verwendet werden. Diese Dateien befinden sich im Hauptordner zusammen mit zwei Unterordnern. Während der Ordner «internal» aus internen Dateien besteht und für die Erstellung selten benutzt wird, befinden sich im Ordner «workspace» viele wichtige Dateien. Darin sind die eigenen Modelle, die Videomaterialien, Datensätze und endgültigen Videoausgaben gespeichert. Beim Installieren der Software werden automatisch zwei Videos mitinstalliert. Diese zwei Videos mit den Namen «data\_dst» und «data\_src» sind die Datensätze, welche für die Erstellung des Deepfakes verwendet werden. Dabei handelt es sich bei der «data\_dst»-Datei um das Zielvideo, auf welchem das Gesicht ausgetauscht wird. Folglich handelt es sich bei der «data\_src»-Datei um die Quelle womit das Zielvideo manipuliert wird. In der vorinstallierten Zielvideodatei sieht man den Unternehmer Elon Musk wie er eine Ansprache hält. Beim Quellvideo handelt es sich um den Schauspieler Robert Downey Junior, ebenfalls in einem Vortrag.

Damit die Software besser verstanden werden konnte, werden die ausführliche Anleitung der grössten Deepfake-Community gelesen sowie nützliche Videos auf YouTube gesucht. Zusammen mit dem Instruktionsvideo von Murtaza Hassan (Hassan, 2020) wird anschliessend ein erster Deepfake erstellt. In dem Video erklärt Hassan in einfachen Worten und verschiedenen Schritten wie mit DFL ein Deepfake erstellt werden kann. Für den Lerneffekt sind die einzelnen Schritte des YouTube-Videos parallel dazu ausgeführt worden. Der vorinstallierte Datensatz mit Elon Musk wird als Zielvideo verwendet. Das Quellvideo ist allerdings mit dem eigenen Videomaterial des Autors ausgetauscht worden. Dabei gilt es zu beachten, dass das Dateiformat «mp4» eingehalten wird. Falls die Datensätze nicht diesem Format entsprechen, können diese einfach im Internet in das richtige Format konvertiert werden. Nach dem Befolgen der einzelnen Vorgänge und eines kleinen Trainings des Modells resultierte daraus bereits der erste Deepfake. Eine Momentaufnahme dieses ersten synthetischen Inhaltes ist in Abbildung 17 sichtbar. Der Deepfake ist nicht perfekt jedoch dauerte die Erstellung dieses kurzen Videos wenige Stunden und lieferte wichtige Erkenntnisse über die Bedienung der Software. Mit der Generierung des ersten eigenen Inhaltes waren die Vorbereitungen abgeschlossen und es konnte damit begonnen werden den geplanten Deepfake zu erstellen.

	Datei	Datei	Datei	Datei
📁	_internal	04.04.2021 17:39	Dateiordner	
📁	workspace	18.05.2021 19:43	Dateiordner	
📄	1) clear workspace.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	2) extract images from video data_src.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	3) cut video (drop video on me).bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	3) extract images from video data_dst FU...	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	3.optional) denoise data_dst images.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	4) data_src faceset extract MANUAL.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	4) data_src faceset extract.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	4.1) data_src view aligned result.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	4.2) data_src sort.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	4.2) data_src util add landmarks debug i...	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	4.2) data_src util faceset enhance.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	4.2) data_src util faceset metadata restore...	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	4.2) data_src util faceset metadata save.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	4.2) data_src util faceset pack.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	4.2) data_src util faceset resize.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	4.2) data_src util faceset unpack.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	4.2) data_src util recover original filenam...	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	5) data_dst faceset extract + manual fix.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	5) data_dst faceset extract MANUAL.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	5) data_dst faceset extract.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	5) data_dst faceset MANUAL RE-EXTRAC...	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	5.1) data_dst view aligned results.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	5.1) data_dst view aligned_debug results....	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	5.2) data_dst sort.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	5.2) data_dst util faceset pack.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	5.2) data_dst util faceset resize.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	5.2) data_dst util faceset unpack.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	5.2) data_dst util recover original filenam...	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	5.XSeg) data_dst mask - edit.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	5.XSeg) data_dst mask - fetch.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	5.XSeg) data_dst mask - remove.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	5.XSeg) data_dst trained mask - apply.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	5.XSeg) data_dst trained mask - remove....	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	5.XSeg) data_src mask - edit.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	5.XSeg) data_src mask - fetch.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	5.XSeg) data_src mask - remove.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	5.XSeg) data_src trained mask - apply.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	5.XSeg) data_src trained mask - remove.b...	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	5.XSeg) train.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	6) train Quick96.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	6) train SAEHD.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	7) merge Quick96.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	7) merge SAEHD.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	8) merged to avi.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	8) merged to mov lossless.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	8) merged to mp4 lossless.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	8) merged to mp4.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	10.misc) make CPU only.bat	04.04.2021 17:39	Windows-Batchda...	1 KB
📄	10.misc) start EBSynth.bat	04.04.2021 17:39	Windows-Batchda...	1 KB

Abbildung 16: Ordner in DFL



Abbildung 17: Erster Deepfake



## 7.2 Erstellung des eigenen Deepfakes

Nachdem die Vorbereitungen und die Grundlagen abgeschlossen sind, beginnt die Phase des eigenen Deepfakes. Als erster Schritt wird die Software erneut heruntergeladen und in einem neuen Ordner auf der externen Festplatte abgespeichert. Anschliessend sind die beiden vorinstallierten Datensätze mit Elon Musk und Robert Downey Junior durch die eigenen Videomaterialien ausgetauscht worden. Dabei müssen die eigenen Datensätze gleich wie die vorinstallierten Dateien benannt und in den «workspace»-Ordner kopiert werden, wobei dadurch die installierten Datensätze ersetzt werden. Wie im Unterkapitel 7.1.3 bereits erwähnt wird, handelt es sich bei der Zieldatei um ein 20 Minuten langes Video. Für einen Deepfake in solchem Ausmass müsste die Trainingszeit des Modells sehr lang sein, was den Rahmen dieser Arbeit überschreiten würde. Ausserdem ist in dem Video die Zielperson nur zur Hälfte der Zeit ersichtlich. Deshalb wird eine geeignete Stelle im Video definiert wobei die Zielperson für 32 Sekunden ohne die Einblendung eines anderen Inhaltes in die Kamera spricht. DFL stellt eine Option zur Verfügung, welche es ermöglicht, ein beliebiges Video schnell auf die gewünschte Länge zu schneiden. Der Inhalt muss dabei auf die bat-Datei mit dem Namen «cut video» gezogen werden. Anschliessend öffnet sich ein Kommandofeld, indem die Start- und Endzeit des gewünschten Videoausschnitts angegeben werden können. Innerhalb weniger Sekunden führt die bat-Datei die Befehle aus und speichert das geschnittene Video im «workspace»-Ordner ab. Darauf muss das bereits bestehende Zielvideo durch den neuen kürzeren Datensatz ersetzt werden. Nachdem dieser Schritt abgeschlossen ist, kann mit der eigentlichen Arbeit für die Erstellung des Deepfakes begonnen werden. Die Herstellung wird in sieben Arbeitspakete aufgeteilt, welche in den anschliessenden Unterkapiteln beschrieben sind.

### 7.2.1 Extrahierung der Quelle

Da es sich bei einem Video im Wesentlichen um eine Sequenz von vielen Bildern handelt, werden im ersten Arbeitsschritt die Bilder des Quellvideos extrahiert. Dafür wird die bat-Datei mit dem Namen «extract images from video data-src» aufgerufen. Mit dieser Funktion werden somit alle Bilder aus dem Video einzeln extrahiert und in einem automatisch erstellten Ordner abgespeichert. Im Kommandofeld gibt es mehrere Optionen für diesen Vorgang. Die Standard-Bildrate des Videos kann verändert werden, indem ein numerischer Wert eingegeben wird. Wenn zum Beispiel der Wert fünf eingegeben wird, ist das Ergebnis, dass fünf Bilder pro Sekunde dargestellt und somit weniger Darstellungen aus

dem Video extrahiert werden. Ebenfalls kann das Format der extrahierten Bilder zwischen «JPG» und «PNG» ausgewählt werden. Da DFL für das Training des Modells nur «JPGs» benutzt, sollte man «PNGs» nicht verwenden. Aus diesem Grund wurde für die Herstellung des eigenen Deepfakes beim Eingabefeld das «JPG»-Format ausgewählt. Die Ausführung dauert nicht länger als eine halbe Minute und aus dem 30 Sekunden-Video werden 767 Bilder in den Ordner mit dem Namen «data\_src» gespeichert, wodurch der erste Arbeitsschritt abgeschlossen ist.

### **7.2.2 Extrahierung des Zielvideos**

Dieser Vorgang funktioniert gleich wie der vorherige. Der einzige Unterschied ist, dass jetzt die Bilder des Zielvideos extrahiert werden. Hierfür wird die bat-Datei «extract images from video data\_dst» angewählt. Beim Kommandofeld stehen wieder die gleichen Optionen zur Verfügung und auch hier werden die Bilder im «JPG»-Format abgespeichert. In kurzer Zeit werden insgesamt 825 Bilder aus dem Video extrahiert und in der richtigen Reihenfolge im automatisch erstellten Ordner «data\_dst» abgespeichert. Somit sind von beiden Datensätzen alle Bilder extrahiert abgespeichert worden.

### **7.2.3 Extrahierung der Gesichter «data\_src»**

Sobald die verschiedenen Sequenzen des Videos als Bilder abgespeichert sind, geht es in diesem Arbeitsschritt um die Extrahierung der Gesichter mit den diversen Emotionen. Für die Ausführung wird die bat-Datei «data\_src faceset extract» gebraucht. Dabei werden die Gesichter unter Verwendung des S3FD-Algorithmus extrahiert (Perov et al., 2020, S. 4). Bei diesem Algorithmus handelt es sich um einen Echtzeit-Gesichtsdetektor, welcher mit einem einzigen neuronalen Netzwerk die Gesichter aus den sequenzierten Bildern extrahiert (Perov et al., 2020, S. 4). Der Algorithmus bietet die Möglichkeit zwischen drei verschiedenen Arten der Gesichtsextrahierung auszuwählen: «full face», «whole face» und «head». Die Standardeinstellung ist «whole face», welche auch für die Erstellung des eigenen Deepfakes verwendet wird. Diese Einstellung wird ebenfalls von der Anleitung empfohlen und gilt für die Extrahierung als universelle Lösung. Als nächstes kann die Auflösung der Bilder bestimmt werden. Hier wird die Standardeinstellung von 128 übernommen. Als letztes erscheint im Kommandofeld die Frage, ob der Prozessor (CPU) oder die Grafikkarte (GPU) für die Extraktion verwendet werden soll. Da die Grafikkarte in der Regel schneller ist, wird diese ausgewählt. Wie bereits erwähnt, ist eine gute Grafikkarte eine der wenigen Voraussetzungen für die Erstellung des eigenen Deepfakes. Während des ganzen Prozesses wird eine «GeForce RTX 20280» Grafikkarte

verwendet. Sobald die Optionen ausgewählt sind, startet der Prozess, wobei nach ungefähr einer Minute die Anzahl der extrahierten Bilder auf dem Display erscheint. Die Anzahl der extrahierten Bilder muss mit der Menge an sequenzierten Bildern übereinstimmen. Falls mehr Gesichtsbilder entdeckt und abgespeichert werden, als sequenzierte Bilder vorhanden sind, muss eine zusätzliche Massnahme ausgeführt werden.

Die extrahierten Gesichter sind im automatisch erstellten Ordner «aligned» abgespeichert worden. Damit die überzähligen Gesichter schnell entdeckt werden können, bietet DFL die bat-Datei mit dem Namen «data\_src view aligned result» an. Es öffnet sich eine externe App, die es ermöglicht die falsch ausgerichteten Quellgesichter zu identifizieren und zu löschen, da diese für den weiteren Prozess nicht gebraucht werden. Weitere bat-Dateien mit verschiedenen Funktionen existieren für diesen Arbeitsschritt. Einige davon werden im nächsten Kapitel erläutert und die restlichen wurden im gesamten Prozess nicht verwendet, sind jedoch in der Deepfake-Anleitung der Community in Anhang A weiter erläutert. Somit ist dieser Ablauf abgeschlossen und es kann mit Arbeitsschritt vier gestartet werden.

#### **7.2.4 Extrahierung der Gesichter «data\_dst»**

Da von Beiden Datensätzen die Gesichtsbilder aus den sequenzierten Bildern für das Training des Modells benötigt werden, müssen auch die Gesichter des Zielvideos extrahiert werden. Der Ablauf ist dabei der Gleiche wie im vorherigen Unterkapitel. Der Unterschied liegt darin, dass die benötigten bat-Dateien einen anderen Namen haben. Inhaltlich sind die gleichen Optionen vorhanden, welche ebenfalls gleich gewählt werden wie bei der Extraktion der Gesichter aus dem Quellvideo. Wie eine solche endgültige Extrahierung aussieht, wird in Darstellung 18 gezeigt. Die Bilder werden im automatisch erstellen Ordner «aligned» abgespeichert.



Abbildung 18: Extrahierte Gesichter

### 7.2.5 Training des Modells

In diesem Arbeitsschritt wird die KI auf die beiden Datensätze trainiert. Für den ersten Deepfake wird die bat-Datei «train Quick96» verwendet. Im Kommandozeilenfenster können als erstes die Modelleinstellungen eingegeben werden. Solange noch kein anderes Modell existiert, muss ein Name für das Modell eingegeben werden. Anschließend erhält man wieder die Aufforderung zur Auswahl zwischen GPU und CPU. Sobald die Datensätze geladen sind, erscheint neben der Kommandozeile das Vorschauenfenster, welches in Abbildung 19 ersichtlich ist. Dabei werden fünf Spalten mit Gesichtern angezeigt. In der ersten Spalte erscheint das Quellgesicht als Beispiel. In der nächsten Kolonne wird das durch DFL gelernte Gesicht angezeigt. Das gleiche gilt für die Spalten drei und vier, wobei dort das Gesicht des Zielvideos ersichtlich ist. Die Kolonne fünf zeigt schliesslich das Finale Gesicht, welches aus dem Training der beiden Datensätzen resultiert. Am Anfang sind die Gesichter der Spalten zwei, vier und fünf sehr verschwommen und ungenau. Dies ändert sich jedoch je länger man das Modell trainiert und die Anzahl der Iterationen ansteigt. Im oberen Bereich des Vorschauenfensters sind die einzelnen Funktionen beschrieben, welche mit verschiedenen Tasten ausgelöst werden, wie zum Beispiel die Aktualisierung der Gesichter mit der Taste «P». Ebenso ist ein Verlustwert-Diagramm vorhanden, welches anzeigt wie gut das Modell trainiert ist. Im ersten Versuch wurde das Modell bis auf 150'000 Iterationen trainiert, wofür DFL etwa vier Stunden brauchte. Wenn die gewünschte Anzahl an Iterationen erreicht wird, kann das Training des Modells beendet und mit dem nächsten Arbeitsschritt begonnen werden.

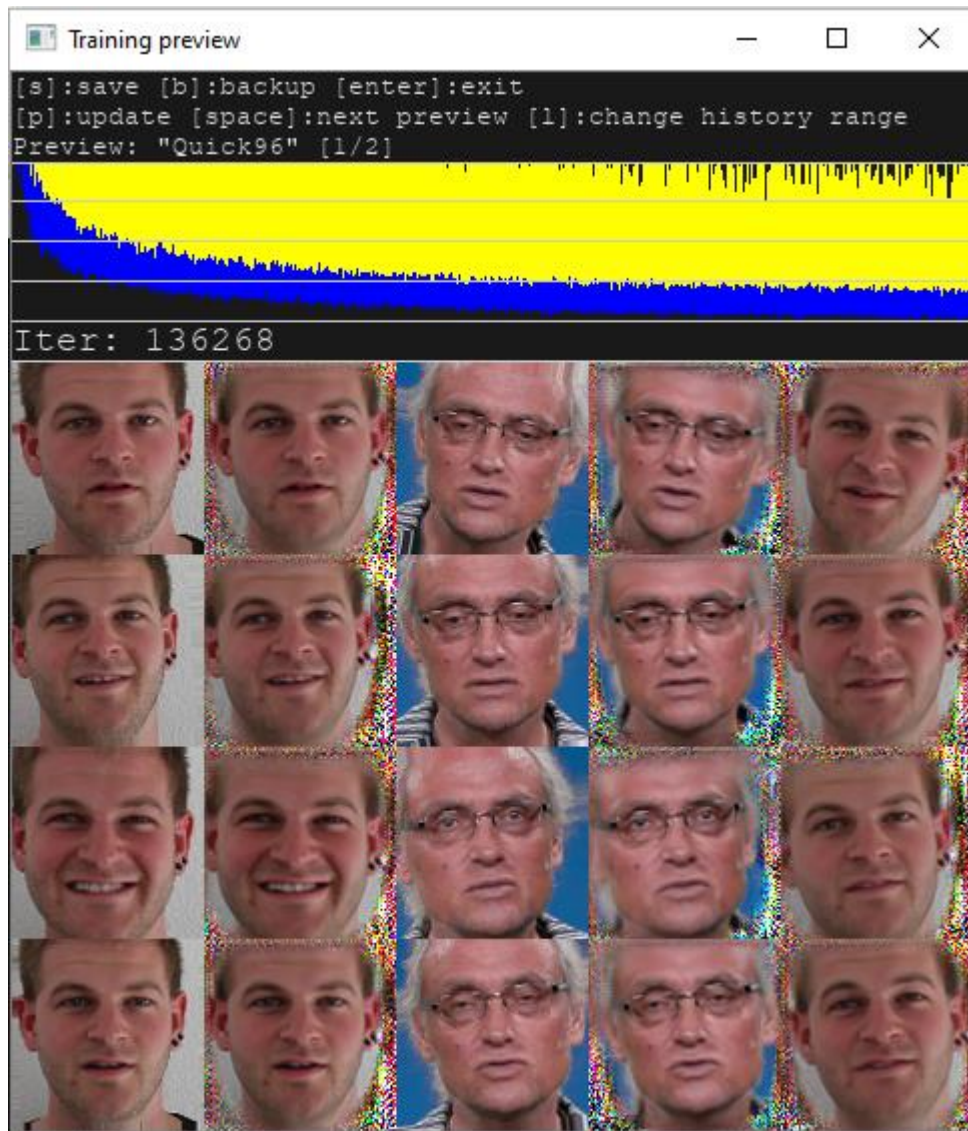


Abbildung 19: Vorschaufenster beim trainieren

### 7.2.6 Zusammenführen

Das trainierte Gesicht muss anschliessend mit dem Zielvideo zusammengeführt werden. Dafür wird die bat-Datei «merge Quick96» verwendet, wobei sich wieder ein Kommandozeilenfenster öffnet und mehrere Eingabeaufforderungen angezeigt werden. Bei der ersten Option geht es darum, ob der Interaktive Konverter verwendet werden soll. In den Standardeinstellungen ist dieser aktiviert, da der Konverter alle Funktionen und eine interaktive Vorschau bietet, in welcher die Auswirkungen aller Änderungen sichtbar sind. Als nächstes muss man das gewünschte Modell für die Zusammenführung auswählen. Wenn bereits mehrere Modelle trainiert wurden, kann eines davon ausgesucht werden, ansonsten wird ein neuer Name eingegeben. Ebenfalls muss die Auswahl zwischen CPU

oder GPU getroffen werden. Wenn die Eingabeaufforderungen ausgefüllt sind, erscheinen unterhalb im Kommandozeilenfenster die aktuellen Einstellungen zum Bild und es öffnet sich parallel dazu ein erneutes Vorschaufenster. Im Vorschaufenster werden als erstes alle Bedienelemente zur Ausführung des interaktiven Konverters angezeigt, welche in Abbildung 20 aufgeführt sind.

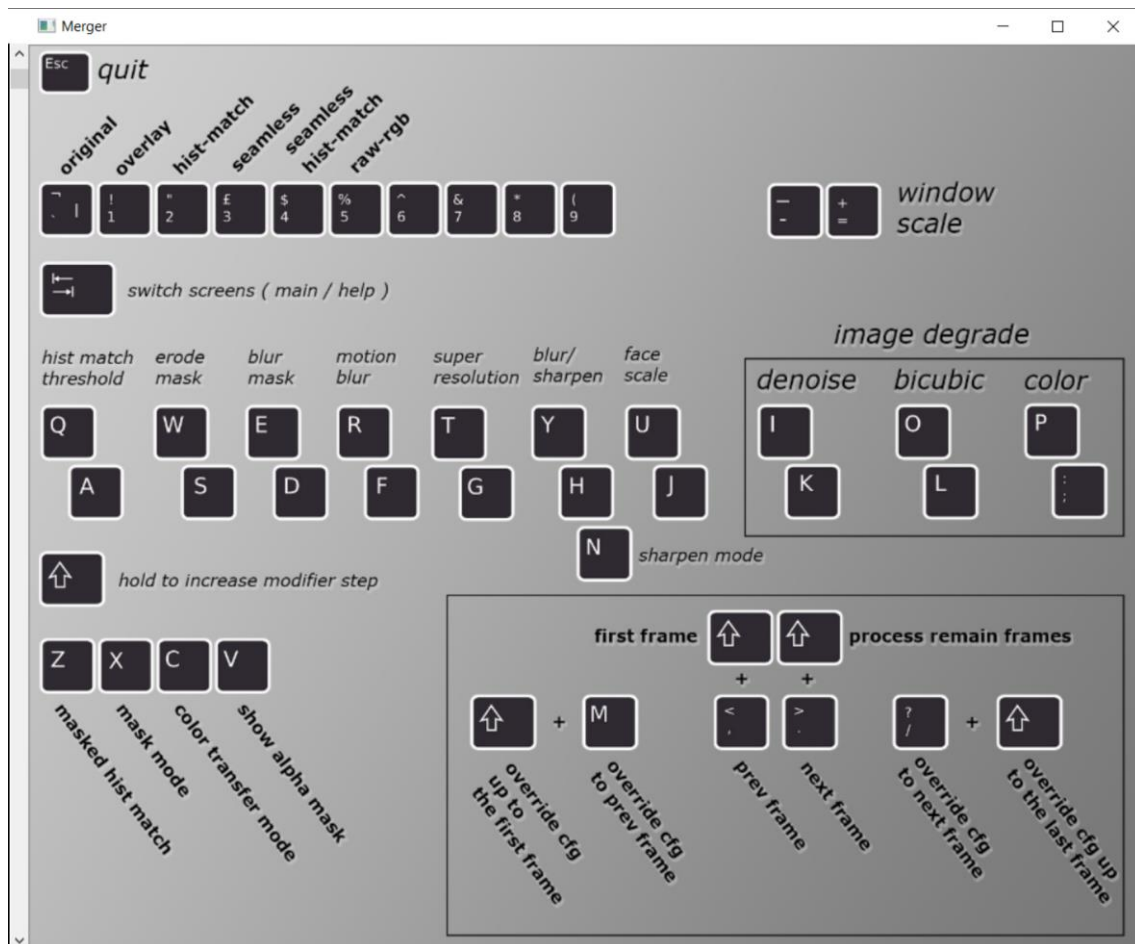


Abbildung 20: Bedienelemente in DFL

Durch die Eingabe der Tabulatortaste ändert das Vorschaufenster von der Übersicht der Bedienelemente zum ersten Bild des Videos. Dabei erscheint die durch das Training generierte Nachbildung des Quellgesichts an der Stelle des Gesichts der Person im Zielvideo. Anschliessend beginnt die grösste Arbeit für die Erstellung des Deepfakes, indem man versucht wird das generierte Gesicht möglichst echt, anstelle des Gesichts der Person im Video, zu platzieren. Dafür können die in Abbildung 20 aufgeführten Optionen und Bedienelemente benutzt werden. Für einen überzeugenden Deepfake müssen die verschiedenen Effekten von DFL ausprobiert werden. Eine detaillierte Beschreibung jedes einzelnen Bedienelements ist unter Anhang A zu finden. Für die Erstellung des ersten

Deepfakes werden die meisten Parameter bei der Schärfe der Maske verändert. Durch das Anklicken einer bestimmten Taste kann zum nächsten Bild gewechselt werden, wobei die vorherigen Einstellungen automatisch abgespeichert und für das nächste Bild übernommen werden. Dadurch ermöglicht DFL ein schnelles Wechseln zwischen den einzelnen Bildern. Die Auflösung sowie die Grösse des generierten Gesichtes werden verändert. Für die Erstellung ist kein bestimmtes Vorgehen ausgewählt worden. Das Ziel ist eine möglichst Echte Übereinstimmung zu erzielen. Sobald die Einstellungen für passend empfunden werden, kann nochmals durch alle einzelnen Szenen durchgeklickt werden. Anschliessend sind alle Einstellungen gespeichert und die beiden Fenster schliessen sich.

### **7.2.7 Video erstellen**

Nachdem alle Gesichter zusammengeführt und die Parameter entsprechend verändert sind, können die einzelnen Bilder in ein Video umgewandelt werden. Für dieses Vorgehen wird die bat-Datei «merged to mp4» verwendet. Diese Datei kombiniert innerhalb von wenigen Sekunden alle Einzelbilder in das gewünschte Format zu einem Video. Wenn DFL diesen Schritt erledigt hat, ist eine Datei mit dem Namen «result» erstellt und automatisch im Ordner «workspace» abgespeichert worden. Diese Datei stellt den ersten eigenen Deepfake dar. Eine Momentaufnahme des Videos ist in Abbildung 21 ersichtlich.



Abbildung 21: Eigener Deepfake

### 7.2.8 Erkenntnisse

Der Prozess zur Erstellung des ersten eigenen Deepfakes und das Video als Ergebnis liefern einige Erkenntnisse. Die gesamte Erstellung des Deepfakes dauerte ungefähr sieben Stunden, wobei davon etwa drei Stunden aktiv am Deepfake gearbeitet wurde. Während der meisten Zeit wird das Modell trainiert, wodurch die Benutzer und Benutzerinnen von DFL nicht aktiv daran arbeiten. Die Ergebnisse der zusammengeführten Gesichter im Video präsentieren, im Vergleich zum damit verbundenen Aufwand, einen glaubhaften Deepfake. Aus dem Video lassen sich einige Erkenntnisse ableiten. Bei einer Frontalaufnahme funktioniert die Software sehr gut. Besonders die Bereiche um die Region der Nase, Mund und Augen werden dabei überzeugend abgebildet und stimmen mit den Emotionen der ursprünglichen Person überein. Ebenso kann durch die Veränderung der Parameter der Unterschied der verschiedenen Hauttypen auf ein Niveau angepasst werden, wodurch dieser Unterschied ohne genauere Betrachtung nicht auffällt. Die Synchronisation mit dem gesprochenen der Person im Zielvideo und den Lippen des generierten Gesichts ist ebenfalls sehr glaubwürdig. Jedoch gibt es in dem kurzen Film auch einige Punkte, welche an die erwähnten Schwachstellen und Herausforderung in Kapitel 3.3 erinnern. Da die Person im Zielvideo eine Brille trägt, ist während des gesamten Videos eine Okklusion vorhanden. Obwohl die Brillenbügel bei den Ohren sowie der äussere Brillenrahmen rund um die Augen zu sehen sind, fehlen im gesamten Video die Brillengläser sowie der Nasenbügel der Brille. Ebenfalls gib es im Video eine kurze Szene, wo die Zielperson nach unten schaut. Dabei entsteht ein verschwommenes Gesicht, wodurch ein Teil des ursprünglichen Gesichts und ein Brillenglas ersichtlich sind. Dieses Problem ist in Abbildung 22 sichtbar. Für weitere Erkenntnisse werden deshalb im nächsten Kapitel diese beiden Herausforderungen genauer behandelt.





Abbildung 22: Posenschwankung und Okklusion

### **7.3 Bearbeitung Okklusion und Posenschwankung**

Der bisherige erstellte Deepfake liefert überzeugende Ergebnisse, jedoch sind in dem kurzen Video zwei Grundlegende Herausforderungen von Deepfakes aufgetreten. Darum werden diese Herausforderungen genauer angeschaut und es wird versucht diese beiden Merkmale zu minimieren, damit der Deepfake einen überzeugenderen Eindruck macht. Dabei werden drei verschiedene Ansätze gewählt, welche in diesem Kapitel beschrieben sind.

#### **7.3.1 Mehr Iterationen**

Für diesen Ansatz wird die Software als erstes erneut installiert und unter einem neuen Namen auf der externen Festplatte abgespeichert. Danach beginnt der im vorherigen Kapitel beschriebene Prozess zur Erstellung des Deepfakes erneut. An den Grundeinstellungen wird nichts verändert, auch werden dieselben Videos als Datensätze verwendet. Der grösste Unterschied zum ersten Deepfake liegt darin, dass die Dauer für das Training verlängert und somit mehr Iteration erzielt werden. Insgesamt wurde das Modell sieben Stunden trainiert und erzielt über 400'000 Iterationen.

Dabei werden die Gesichter im Vorschaufenster detaillierter. Anschliessend werden die Parameter beim Zusammenführen der Gesichter neu eingestellt. Da die Einstellungen bei der Frontalaufnahme bereits gute Ergebnisse erzielt haben, sind dort jeweils kleine Verbesserungen vorgenommen worden. Beim ersten Deepfake werden die veränderten Einstellungen für alle sequenzierten Bilder übernommen. In diesem Versuch sind die Werte der Szene mit der Kopfsenkung gegenüber den Frontalaufnahmen jedoch unterschiedlich. Die Schärfe des Bildes und die Breite des generierten Gesichtes sowie weitere kleine Einstellungen sind verändert worden.

Obwohl das Modell deutlich länger trainiert wurde und somit das generierte Gesicht bei den Frontalaufnahmen echter erscheint, hatte die längere Trainingsdauer auf die beiden Herausforderungen wenig Auswirkungen. Die getätigten Veränderungen bei den Parametern haben zur Folge, dass beim Abspielen des Videos, Unterschiede zwischen den einzelnen Szenen zu erkennen sind. Somit ist ersichtlich, dass das Gesicht und die Schärfe im Vergleich zum Rest des Videos verändert worden sind. Ebenfalls hatten die Einstellungen nur einen minimalen Verbesserungseffekt im Vergleich zum ersten Deepfake zur Folge. Dabei kann die Erkenntnis gewonnen werden, dass die Quelldatei zu wenig differenzierbare Kopfposen enthält. Dadurch kann die KI nicht auf eine neigende Kopfhaltung trainiert werden, weshalb trotz der Steigerung der Iterationen kein deutlich besseres Ergebnis erzielt wird.

### **7.3.2 Neues Videomaterial**

Trotz wenigen positiven Veränderungen unter Verwendung des ersten Ansatzes, können davon eine wichtige Erkenntnis erschlossen werden. Das bisherige Videomaterial der Quelle besteht nur aus einer Frontalaufnahme mit verschiedenen Emotionen. Darin fehlen jedoch Szenen, bei welchen die Person im Video nach links, rechts sowie nach unten schaut. Dadurch musste das bisherige Modell lediglich die Vorderseite des Gesichtes erlernen und kann bei einer Neigung des Kopfes kein passendes Bild zuordnen, was wiederum zu einer Verschlechterung der Qualität des Deepfakes führt. Aus diesem Grund wird bei diesem Ansatz mit einem neu aufgenommenen Video des Autors gearbeitet. Beim Video wurde diesmal darauf geachtet, dass mehrere Aufnahmen des Gesichtes von allen Seiten existieren. Das neue Video wird ebenfalls auf der externen Festplatte abgespeichert und gleichzeitig wird DFL erneut heruntergeladen und in einem neuen Ordner eröffnet.

Die einzelnen Arbeitsschritte sind die Gleichen wie bisher. Neben dem neuen Videomaterial wird bei diesem Versuch das Modell möglichst lange trainiert und erreicht ebenfalls über 400'000 Iterationen. Im Unterschied zu den vorherigen Versuchen wird im Vorschaufenster ersichtlich, dass auch die gewünschten links, rechts Aufnahmen sowie die Neigung nach unten, trainiert werden.

Bei der folgenden Einrichtung der Parameter werden für die Frontalaufnahmen dieselben Einstellungen getätigt wie beim vorherigen Versuch, da mit diesen ein überzeugendes Ergebnis erzielt wurde. Bei der Drehung des Kopfes nach rechts, konnten mit Hilfe der neuen Bilder und dem Verändern der Einstellungen einige Verbesserungen erzielt werden. Der Übergang des generierten Gesichtes zum Rest des Kopfes der Zielperson ist weniger sichtbar. Auch ist die Bewegung weniger unscharf als in den vorherigen Versuchen. Ein Problem bei dieser Gesichtshaltung bleibt weiterhin die Brille. Während beim rechten Auge die Brille deutlich zu sehen ist, fehlen der Nasenbügel sowie das Brillenglas und der Brillenrand beim linken Auge komplett. Diese Änderung der Kopfhaltung ist zwar eine sehr kurze Szene im Video, jedoch wird der Deepfake dadurch unglaubhafter. Bei der Neigung des Kopfes konnten ebenfalls einige positive Ergebnisse erzielt werden. Im Gegensatz zu den bisherigen Versuchen verschwindet das synthetische Gesicht nicht vollständig, sondern bleibt vor allem auf der rechten Seite gut erhalten. Der gesamte Bereich ist nicht mehr komplett verschwommen und enthält klare Gesichtszüge. Ein Problembereich bildet weiterhin die linke Seite des generierten Gesichtes, welches verschwommen und durchsichtig ist. Beide Szenen sind in einer Momentaufnahme in Abbildung 23 ersichtlich, wobei die Problemzonen rot umrandet sind.



Abbildung 23: Ergebnisse mit neuem Videomaterial

Der Ansatz mit dem neuen Videomaterial hat einige Verbesserungen erbracht und bei einer nicht allzu genauen Betrachtung des Deepfakes kann das Video als nicht manipulierter Inhalt betrachtet werden. Damit war die im vorherigen Ansatz erschlossene Erkenntnis korrekt und hat zu einem besseren Ergebnis geführt. Jedoch sind die Problemzonen im Video noch nicht perfekt und müssen mit weiteren Ansätzen betrachtet werden. DFL bietet nebst der «train Quick96»-Datei eine weitere Funktion mit dem Namen «train SAEHD». Diese Ausführung bietet mehr Einstellungsmöglichkeiten für das Training des Modells und wird deshalb im nächsten Ansatz angewendet.

### 7.3.3 Training mit SAEHD

Bei diesem Ansatz werden nebst der neuen Auswahl für das Trainieren des Modells noch weitere Aspekte geändert. Bei den bisherigen Versuchen lag der Fokus auf den Problemzonen, jedoch wurden auch die restlichen Bildsequenzen mitberücksichtigt, was vor allem beim Training mehr Zeit in Anspruch genommen hat. Deshalb wird das Zielvideo bei diesem Ansatz zuerst geschnitten und von 30 auf 15 Sekunden reduziert. Dadurch liegt der Schwerpunkt noch mehr auf den Szenen mit den Herausforderungen und beim Modell werden die verschiedenen Gesichtsposen verstärkter trainiert. Das Schneiden des Videos wird erneut durch die bat-Datei «cut Video» durchgeführt. Die Literaturrecherche im Kapitel 2 zeigte zudem, dass Nachahmungen bei Deepfakes zu einem besseren Ergebnis

führen. Darum wird bei diesem Ansatz ein neues Quellvideo eingesetzt, wobei der Autor die Problemszenen nachspielt, und versucht die gleichen Kopfbewegungen wie die Person im Zielvideo zu machen. Dieses Video ist ebenfalls kürzer als die bisherigen Quellvideos. Anschliessend werden beide Datensätze in einem neu angelegten Ordner abgespeichert und die ersten Arbeitsschritte werden bis auf eine Ausnahme wie bisher durchgeführt. DFL bietet eine Funktion an, bei welcher ein spezieller Algorithmus des maschinellen Lernens das Aussehen der Gesichter im Datensatz des Quellvideos verbessert. Dies kann nützlich sein, wenn der Datensatz ein wenig unscharf ist oder wenn die Bilder noch detaillierter werden sollen. Die Funktion wurde bisher nicht verwendet, damit ein Unterschied zu den bisherigen Ergebnissen festgestellt werden kann. Auch dieser Prozess dauert nicht länger als eine Minute und die Bilder werden automatisch im Ordner abgespeichert. Danach werden die Arbeitsschritte bis zum Training des Modells wie bislang durchlaufen.

Es wurde bereits erläutert, dass für das Trainieren des Modells zwei bat-Dateien zur Verfügung stehen, wobei jetzt nur eine davon ausgeführt wurde. In diesem Ansatz wird jedoch die noch nicht verwendete Funktion für den Versuch ausgewählt. Dabei können in der «train SAEHD»-Datei viel mehr Optionen betreffend dem Training eingestellt werden. Die verschiedenen Möglichkeiten erscheinen wieder im öffnenden Kommandozeilenfenster und dauern insgesamt länger als der bisherige Prozess bis zu diesem Arbeitsschritt. Die Beschreibungen zu den einzelnen Einstellungen sind ebenfalls in der Anleitung der Community in Anhang A zu finden. Dabei handelt es sich um Optionen wie die Farbübertragung, die Auflösung, Zielanzahl Iterationen und Spiegelung der Flächen, um einige Beispiele zu nennen. Da beim Modell viel mehr zusätzliche Einstellungen aktiviert werden, dauert das Üben deutlich länger. Aus diesem Grund wurde das Modell etwa 7 Stunden lang trainiert und erreichte dabei nur über 100'000 Iterationen. Aus zeitlichen Motiven wurde das Modell nicht länger trainiert.

Beim Zusammenführen der Gesichter werden bei den Frontalaufnahmen die gleichen Einstellungen verwendet wie bisher. Dabei sind vor allem die Bereiche der Backen sowie des Bartes detaillierter dargestellt als bei den vorherigen Ansätzen. In Abbildung 23 sind die Szenen mit den Herausforderungen erneut dargestellt, wobei die erzielten Fortschritte mit grün markiert werden. Durch den zusätzlichen Einsatz der Verbesserungsfunktion sind Fortschritte im Bereich der Detailierung erzielt worden. Jedoch ist es weiterhin nicht

möglich die Gesamte Brille darzustellen. Bei der Neigung des Kopfes erscheinen vor allem die Mundpartie sowie das rechte Auge ausführlicher. Die linke Seite des Quellgesichtes ist trotzdem zu gering ersichtlich und einige verschwommenen Elemente sind weiterhin sichtbar.



Abbildung 24: Ergebnisse mit SAEHD

#### **7.4 Zusammenfassung eigener Deepfake**

In diesem Kapitel sind der Ablauf zur Erstellung des eigenen Deepfakes erläutert sowie drei verschiedene Ansätze zur Verbesserung von zwei aufgetretenen Herausforderungen präsentiert worden. Durch diesen Versuch ist die Aussage betreffend der einfachen Zugänglichkeit zu Anwendungen für Deepfakes bestätigt worden. Im Internet existiert eine grosse Anzahl von Anleitungen zu den verschiedenen Tools, womit Neulinge schnell mit den Funktionen vertraut werden. Die Erstellung eines Deepfakes, welcher aus einem Video mit reinen Frontalaufnahmen besteht, ist zudem ziemlich einfach und es wird innert weniger Stunden ein glaubwürdiger synthetischer Inhalt produziert. Dabei hängt das Ergebnis oft von den ausgewählten Datensätzen und der investierten Zeit für das Trainieren des Modells ab. Je mehr sich die beiden Datensätze in den verschiedenen Posen und Gesichtsausdrücken ähneln, desto glaubwürdiger wird das Endergebnis. Den grössten Einfluss auf das Resultat hat das Trainieren des Modells. Durch ein längeres Training wird das erzeugte Gesicht detaillierter und die Gesichtszüge sind markanter. Daher kann bei der Intensivierung dieser beiden Arbeitsschritte ein glaubwürdiger Deepfake erzeugt werden.

Die drei Ansätze zur Verbesserung der Okklusion und der Posenschwankung im Zielvideo haben die gewünschten Veränderungen im Endergebnis nicht komplett erfüllt. Trotzdem können davon einige Erkenntnisse gewonnen werden. Wie bereits im vorherigen Abschnitt erwähnt, kann davon ausgegangen werden, dass die Fehler in der Posenschwankung durch ein intensiveres Training sowie einer besseren Nachstellung der Szene im Quellvideo verbessert werden. Es existieren bereits einige Deepfakes mit ähnlichen Posen, wobei die Inhalte überzeugender und echter erscheinen. DFL bietet ausserdem noch weitere Funktionen an, welche solche Herausforderung überwinden. Diese haben jedoch den Aufwand dieser Arbeit überschritten und können daher nicht ausgeführt werden. Ein grösseres Problem stellt die Okklusion, in diesem Fall die Brille der Person im Zielvideo, dar. Bisher wurden in der Deepfake Community noch keine Ansätze entwickelt, welche mit Okklusionen umgehen können. Auch wenn die Brillenträger bei der Frontalaufnahme verschwinden, tauchen die Brillengläser bei der Posenschwankung wieder auf. Für die Erstellung von glaubhaften Deepfakes zu gutartigen Zwecken müssen Methoden erforscht werden, welche Bilder mit Okklusionen generieren können. Da KI noch eine neue Technologie ist, steigt die Wahrscheinlichkeit, dass Wissenschaftler und Deepfake-Communities in den nächsten Jahren in der Lage sind, einen Deepfake mitsamt Okklusionen zu erstellen.

## **8 Schlussteil**

In diesem letzten Kapitel werden die Arbeit und die darin gewonnenen Erkenntnisse reflektiert und kritisch hinterfragt. Als erstes wird die Forschungsfrage beantwortet und anschliessend die Ergebnisse der Hypothesen erläutert. Des Weiteren wird ein Fazit formuliert.

### **8.1 Beantwortung der Forschungsfrage**

Die für diese Bachelorarbeit formulierte Forschungsfrage lautet «Wie ist der Forschungsstand zur Verbreitung und Erstellung von Deepfakes?». Die Beantwortung der Forschungsfrage wird in diesem Kapitel in die zwei Themengebiete der Verbreitung sowie der Erstellung von Deepfakes unterteilt.

Die Literaturrecherche hat gezeigt, dass aktuell wenig wissenschaftliche Arbeiten zur aktuellen Verbreitung von Deepfakes existieren. Die präsentierten Werte in Kapitel 3.1 weisen auf einen starken Anstieg in den letzten drei Jahren hin, wobei sich die Anzahl von

Deepfakes alle sechs Monate verdoppelt. Durch die analysierten Deepfakes sind ein enormer Unterschied der Geschlechter festgestellt und ein genaues Abbild der Berufe der Opfer aufgezeigt worden. Anhand der Diversität der Nationalitäten von den Opfern lassen sich Deepfakes zudem als globales Phänomen beschreiben. Zu den grössten Verbreitungskanälen von Deepfakes gehören Soziale-Medien wie YouTube, Facebook und TikTok sowie verschiedene Deepfake-Pornowebseiten. Da vor allem Social-Media-Plattformen eine zunehmend zentrale Rolle bei der Verteilung von Informationen für die Öffentlichkeit spielen, wird die Geschwindigkeit der Verbreitung von Deepfakes in Zukunft noch mehr ansteigen. Der unregulierte Charakter von diesen Plattformen und die stetigen Verbesserungen der synthetischen Inhalte werden die Bestimmung aktueller Zahlen zur Verbreitung erschweren. Nebst einigen Anwendungen von privaten Unternehmen existieren zum jetzigen Zeitpunkt keine einheitlichen Methoden zur Überwachung und Verbreitung von Deepfakes.

Für die Erstellung von Deepfakes existieren mehrere wissenschaftliche Arbeiten, welche den aktuellen Forschungsstand erläutern. Besonders die Forschungen von Mirsky und Lee (2021), Masood et al. (2021) und Juefei-Xu et al. (2021) zeigen den aktuellen Stand zur Erstellung von Deepfakes auf. Aktuell können Deepfakes in die unter Kapitel 3.2 aufgeführten fünf verschiedenen Methoden unterteilt werden. Dabei entstehen seitens der Forschung stetig neue Generierungsmethoden. Mittlerweile wird KI in Apps wie Reface oder Fake-App integriert, womit innerhalb von wenigen Sekunden ein kurzer Deepfake zur Unterhaltung erstellt wird. Die Generierung eines eigenen Deepfakes hat gezeigt, dass die Zugänglichkeit und die einzelnen Arbeitsschritte ohne Expertenkenntnisse durchgeführt werden können. Durch diverse Anleitungen auf YouTube und in den einzelnen Web-Foren wird das Wissen für Neulinge bereitgestellt. Die neue Technologie hat eine globale Anhängerschaft, welche sich in organisierten Web-Communities auf Reddit, Github und Deepfake-Pornoplattformen austauschen und ebenfalls neue Methoden und Wege entwickeln, um synthetische Inhalte überzeugender wirken zu lassen. Die fünf Erzeugungsmethoden von Deepfakes werden in Zukunft miteinander kombiniert, wodurch der perfekte Deepfake entstehen kann.



## 8.2 Beurteilung der Hypothesen

In diesem Abschnitt werden die in der Arbeit erzielten Erkenntnisse und Ergebnisse in Bezug auf die formulierten Hypothesen beurteilt.

### 8.2.1 Hypothese 1

H1: Seit der Entdeckung von Deepfakes sind verschiedene Generierungsmethoden entwickelt worden.

Bereits einige Monate nachdem der erste Deepfake auf Reddit veröffentlicht und der dazugehörige Programmcode geteilt wurde, entstanden Verbesserungen und Anpassungen im Code. In der Zwischenzeit entwickelten Deepfake-Communities und Forschende neue Methoden und Anwendungen zur Generierung der synthetischen Inhalte. Die Literaturrecherche hat, die in Kapitel 3.2 aufgeführten fünf Kategorien zur Erzeugung von Deepfakes identifiziert:

- Face-Swap
- Lip-Syncing (Lippensynchronisation)
- Puppet-mastery (Nachahmung)
- Audio-Deepfakes
- Gesichtssynthese und Merkmalmanipulation.

Aktuell konzentriert sich die Generierung von Deepfakes hauptsächlich auf den Gesichtsbereich, wobei die nächste Generation von Deepfakes auch Ganzkörpermanipulationen ermöglichen soll. Dadurch können zum Beispiel die Körperhaltung verändert und mit überzeugenden Gesichtsm Manipulationen kombiniert werden. Ausserdem können Deepfakes bereits heute durch die Bearbeitung von Gesichtsinhalten und synthetisch hergestelltem Audio kombiniert werden. Dadurch, dass seit der Entdeckung der neuen Technologie stets neue Anwendungen sowie bisher fünf verschiedene Erzeugungsmethoden entstanden sind, kann Hypothese 1 **verifiziert** werden.

### 8.2.2 Hypothese 2

H2: In der Schweizer Gesellschaft herrscht ein zu geringes Verständnis über die neue Technologie.

Für die Wiederlegung dieser Hypothese wurden bei insgesamt elf Zeitungen aus mehreren Landesregionen der Schweiz, die Artikel nach dem Wort Deepfake durchsucht. Dabei wurde der Zeitraum von Dezember 2017 bis April 2021 betrachtet. Die in Kapitel 3.2 erwähnten Ergebnisse ergaben eine Gesamtauswertung von 115 Publikationen. Die Qualität und Tiefe der Artikel variiert stark, da das Thema in einem grossen Teil der Artikel nur kurz erwähnt wird. Wenige der identifizierten Artikel liefern einen aufklärenden Überblick über die Technologie, wobei nebst den Bedrohungen auch die positive Nutzung beschrieben werden. Gemäss Schick (2020, S. 261) ist es wichtig, dass beide Seiten der Medaille beschrieben werden, da die Technologie für sich weder gut noch böse ist und durch eine Intensivierung der Aufklärung das Verständnis in der Bevölkerung gestärkt werden kann. Da Deepfakes in der Entwicklung immer noch am Anfang sind, kann durch Informations- und Analysepublikationen in den Massenmedien gute Präventionsarbeit geleistet werden. Die im Rahmen dieser Arbeit durchgeführte Recherche hat allerdings wenig solche Artikel in der Zeitungslandschaft der Schweiz entdeckt. Wegen der überschaubaren und groben Berichterstattung in der Schweiz zum Thema Deepfake, kann die Annahme somit **verifiziert** werden.

Die Verifizierung ist unter Vorbehalt zu betrachten, da die Recherche nicht die gesamte Medienlandschaft der Schweiz abdeckte und keine Umfrage stattgefunden hat. Die Annahme aus Hypothese 2 könnte mithilfe einer Umfrage zum aktuellen Wissensstand über Deepfakes, bekräftigt werden.

### 8.2.3 Hypothese 3

H3: Das Verhältnis zwischen den Nutzen und Bedrohungen durch Deepfakes ist nicht im Gleichgewicht.

Wie in Kapitel 4 beschrieben wird, können Deepfakes für gute oder bösartige Zwecke eingesetzt werden. Obwohl die Technologie ein grosses Potenzial aufweist und in einigen Fällen bereits angewendet wurden, finden Deepfakes heutzutage vor allem in Deepfake-

Pornos ihre Anwendung. Während in der Anfangszeit vor allem berühmte weibliche Persönlichkeiten betroffen waren, sind durch Anwendungen wie DeepNude mittlerweile alle weiblichen Personen betroffen. Nebst dieser Beschämenden Anwendung können Deepfakes vor allem im Bereich der Politik grossen Schaden anrichten. Synthetische Inhalte von politischen Persönlichkeiten können das Denken der Gesellschaft beeinflussen sowie deren Gegner zu einer Gegenaktion veranlassen. Auf der einen Seite kann durch die neue Technologie Menschen geholfen werden, wie beim Projekt Revoice, oder sie versetzt die Betrachter ins Staunen und ist amüsierend. Im Gegenteil dazu kann sie jedoch auch das Leben von einigen Personen innerhalb von wenigen Stunden zerstören, indem etliche synthetische Pornoinhalte erstellt werden. Ebenfalls werden Deepfakes bereits heutzutage manipulativ von verschiedenen Regierungen und Organisationen eingesetzt, wodurch gesellschaftliche Ansichten geändert und aufgehetzt werden können. Im schlimmsten Falle führt ein Deepfake zum Krieg von verschiedenen ethnischen Gruppierungen oder Nationen, indem die Aussagen einer bedeutenden politischen Person abgeändert werden. Aufgrund der genannten Gründe und der Literaturrecherche kann Hypothese 3 **verifiziert** werden.

#### **8.2.4 Hypothese 4**

H4: Für die Bekämpfung von gefährlichen Deepfakes existieren einsatzfähige Lösungsansätze.

Die Wissenschaft fokussierte sich in den vergangenen Jahren vor allem darauf, Methoden zur Erkennung von Deepfakes zu entwickeln, indem KI mit den eigenen Mitteln bekämpft wird. Dabei gelangen der Wissenschaft einige Durchbrüche, welche in Kapitel 6.1 erläutert werden. Trotz einer intensiven Forschung sind bisher keine Tools für den täglichen Einsatz entwickelt worden. Es existieren einige Firmen, welche ihre Dienste im Namen der Cyber-Security gegen die Bedrohung durch Deepfakes anbieten. Bisher sind die entwickelten Methoden der Wissenschaft jedoch nicht in freier Wildbahn zum Einsatz gekommen, da gewisse Kriterien nicht erfüllt werden. Deepfakes können allerdings nicht nur mit technischen Tools bekämpft werden, es braucht auch rechtliche Schritte seitens des Staates. In den Jahren 2019 und 2020 sind in wenigen Ländern die ersten Gesetze gegen synthetische Inhalte in Kraft getreten. Obwohl diese gesetzlichen Richtlinien ein wichtiger Schritt in die richtige Richtung sind, müssen sie auch kritisch hinterfragt werden. Bei der Formulierung der Gesetze ist es wichtig, dass nur die böswillige Nutzung von Deepfakes bestraft wird und die Technologie nicht von Grund auf verboten wird. Die

Niederschrift von solchen Gesetzen und deren Einhaltung stellen die staatlichen Gesetzgeber vor eine Herausforderung, welche bisher wenige Nationen in Angriff genommen haben. Die beste Möglichkeit zur Aufdeckung von bösartigen Deepfakes und der Förderung von gutartigen synthetischen Inhalten besteht in der Kombination von technischen Tools, rechtlichen Grundlangen und im Bewusstsein der Gesellschaft über die Technologie. Die Bearbeitung der Literatur hat ergeben, dass eine solche Kombination weltweit nicht existiert oder in einem solchen Umfang eingesetzt wird, weshalb Hypothese 4 als **widerlegt** beurteilt wird.

### **8.3 Fazit**

In dieser Bachelorarbeit wurde der Forschungsstand zu Verbreitung und Erstellung von Deepfakes untersucht. Die Recherche hat ein grosses Forschungspotential im Bereich der Verbreitung von Deepfakes ergeben, da aktuell nur eine geringe Anzahl wissenschaftlicher Arbeiten zu diesem Thema existieren. Damit vor allem böswillige Deepfakes in Zukunft besser entdeckt werden können, ist es wichtig einen Überblick über deren Verbreitungslandschaft zu haben. In den letzten Jahren sind fünf Erstellungsmethoden für Deepfakes entstanden, welche in Unterschiedlichen Bereichen eingesetzt werden. Aufgrund des immensen technologischen Fortschritts können Deepfakes schnell und einfach ohne Expertenwissen erstellt werden. Die meisten Deepfakes werden heutzutage für bösartige Zwecke gegen Einzelpersonen, Firmen oder Staaten eingesetzt und sind dafür eine Gefahr für die gesamte Gesellschaft. Obwohl Deepfakes ein enormes Potential bieten, werden die Nutzen der Technologie noch zu wenig genutzt. Zur Verteidigung gegen gefährliche synthetische Inhalte muss eine Kombination von technischen Tools, rechtlichen Grundlagen und allgemeinem Verständnis in der Bevölkerung eingesetzt werden, ohne dass Deepfakes verboten werden. Die Erstellung eines eigenen Deepfakes hat weitere Einblicke in den Erstellungsprozess von einem synthetischen Video ergeben. Die gewonnenen Erkenntnisse und Herausforderungen sind kritisch analysiert und diskutiert worden.

## Literaturverzeichnis

- Agarwal, S., Farid, H., Fried, O., & Agrawala, M. (2020). Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2814–2822. <https://doi.org/10.1109/CVPRW50498.2020.00338>
- Ajder, H., Patrini, G., & Cavalli, F. (2020). *Automated Image Abuse—Deepfake Bots on Telegram*. Abgerufen von: <https://sensity.ai/reports/>
- Ajder, H., Patrini, G., Cavalli, F., & Cullen, L. (2019). *Deeprtrace—The State of Deep-fakes*. Abgerufen von: <https://sensity.ai/reports/>
- Albahar, M., & Almalki, J. (2019). Deepfakes: Threats and Countermeasures Systematic Review. *Journal of Theoretical and Applied Information Technology*, 97(22), 3242–3250.
- Ambalina, L. (2020). *How AI is changing the video game industry: Augmentation and synthetic media*. Abgerufen von: [https://www.aibusiness.com/author.asp?section\\_id=789&doc\\_id=761220](https://www.aibusiness.com/author.asp?section_id=789&doc_id=761220)
- Arik, S. O., Chen, J., Peng, K., Ping, W., & Zhou, Y. (2018). Neural Voice Cloning with a Few Samples. *arXiv:1802.06006*. Abgerufen von: <http://arxiv.org/abs/1802.06006>
- Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein GAN. *arXiv:1701.07875*. Abgerufen von: <http://arxiv.org/abs/1701.07875>
- Artbreeder. (2021). Abgerufen von: <https://artbreeder.com>
- Bitouk, D., Kumar, N., Dhillon, S., Belhumeur, P., & Nayar, S. K. (2008). Face swapping: Automatically replacing faces in photographs. *ACM Transactions on Graphics*, 27(3), 1–8. <https://doi.org/10.1145/1360612.1360638>
- BitTorrent. (2021). Abgerufen von: <https://www.bittorrent.com/>

- Blanz, V., Basso, C., Vetter, T., Poggio, T., Brunet, P., & Fellner, D. (2003). Reanimating Faces in Images and Video. *EUROGRAPHICS 2003: the European Association for Computer Graphics, 24th Annual Conference, Blackwell, 641-650 (2003)*. Abgerufen von: [https://www.researchgate.net/publication/47861001\\_Reanimating\\_Faces\\_in\\_Images\\_and\\_Video](https://www.researchgate.net/publication/47861001_Reanimating_Faces_in_Images_and_Video)
- Breeden, A. (2020, Februar 5). Defense Minister Was on the Line, Asking for Millions to Aid France. Or Was He? *The New York Times*. Abgerufen von: <https://www.nytimes.com/2020/02/04/world/europe/france-Jean-Yves-Le-Drian-fraud.html>
- Brown, N. I. (2020). Deepfakes and the Weaponization of Disinformation. *Virginia Journal of Law & Technology Association, 23(01)*.
- Buo, S. A. (2020). *The Emerging Threats of Deepfake Attacks and Countermeasures*. <https://doi.org/10.13140/RG.2.2.23089.81762>
- Caporusso, N. (2021). Deepfakes for the Good: A Beneficial Application of Contentious Artificial Intelligence Technology. In T. Ahram (Hrsg.), *Advances in Artificial Intelligence, Software and Systems Engineering* (S. 235–241). Springer International Publishing. [https://doi.org/10.1007/978-3-030-51328-3\\_33](https://doi.org/10.1007/978-3-030-51328-3_33)
- Cavalli, F. (2021, Februar 8). *How to detect a deepfake online with no coding skills*. Abgerufen von: <https://sensity.ai/how-to-detect-a-deepfake/>
- Chang, Y.-J., & Ezzat, T. (2005). Transferable videorealistic speech animation. *Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation, 143–151*. <https://doi.org/10.1145/1073368.1073388>
- Charles, J., Magee, D., & Hogg, D. (2016). Virtual Immortality: Reanimating Characters from TV Shows. In G. Hua & H. Jégou (Hrsg.), *Computer Vision – ECCV 2016 Workshops* (Bd. 9915, S. 879–886). Springer International Publishing. [https://doi.org/10.1007/978-3-319-49409-8\\_71](https://doi.org/10.1007/978-3-319-49409-8_71)

- Chesney, R., & Citron, D. K. (2018). *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security* (SSRN Scholarly Paper ID 3213954). Social Science Research Network. <https://doi.org/10.2139/ssrn.3213954>
- Cole, S. (2017, Dezember 11). *AI-Assisted Fake Porn Is Here and We're All Fucked*. Abgerufen von: <https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn>
- Cole, S. (2018, Januar 24). *We Are Truly Fucked: Everyone Is Making AI-Generated Fake Porn Now*. Abgerufen von: <https://www.vice.com/en/article/bjye8a/reddit-fake-porn-app-daisy-ridley>
- Davis, R. (2020). *Technology Factsheet: Deepfakes*. Harvard Kennedy School. Abgerufen von: <https://www.belfercenter.org/publication/technology-factsheet-deepfakes>
- DeepFaceLab. (2021, Mai 15). Abgerufen von: <https://github.com/iperov/DeepFaceLab> (Original work published 2018)
- Delfino, R. A. (2020). Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn's Next Tragic Act. *Actual Problems of Economics and Law*, 14(1). <https://doi.org/10.21202/1993-047X.14.2020.1.105-141>
- Deshmukh, A., & Wankhade, S. B. (2021). Deepfake Detection Approaches Using Deep Learning: A Systematic Review. In V. E. Balas, V. B. Semwal, A. Khandare, & M. Patil (Hrsg.), *Intelligent Computing and Networking* (S. 293–302). Springer. [https://doi.org/10.1007/978-981-15-7421-4\\_27](https://doi.org/10.1007/978-981-15-7421-4_27)
- Elevenpaths. (2019). *Artificial Intelligence: GANs and Autoencoders applied to Cyber-Security*. Abgerufen von: <https://pro-cdo-web-resources.s3.eu-west-1.amazonaws.com/elevenpaths/uploads/2020/6/elevenpaths-whitepaper-artificial-intelligence-gans-and-autoencoders-applied-to-cybersecurity.pdf>
- FaceApp. (2021). Abgerufen von: <https://faceapp.com/>

- Fan, B., Wang, L., Soong, F., & Xie, L. (2015, April 19). *Photo-real talking head with deep bidirectional LSTM*. <https://doi.org/10.1109/ICASSP.2015.7178899>
- Farid, H., & Schindler, H.-J. (2020, Juni 28). Deepfakes—Eine Bedrohung für Demokratie und Gesellschaft. *Counter Extremism Project*. Abgerufen von: <https://www.kas.de/de/einzeltitel/-/content/die-gefahr-von-deep-fakes-fuer-unsere-demokratie>
- Ferraro, M. (2019). *Deepfake Legislation: A Nationwide Survey*. Abgerufen von: <https://www.wilmerhale.com/en/insights/client-alerts/20190925-deepfake-legislation-a-nationwide-survey>
- Fleishman, G. (2019, April 30). How to spot the realistic fake people creeping into your timelines. *Fast Company*. Abgerufen von: <https://www.fastcompany.com/90332538/how-to-spot-the-creepy-fake-faces-who-may-be-lurking-in-your-timelines-deepfaces>
- Gardiner, N. (2019). *Facial re-enactment, speech synthesis and the rise of the Deepfake* [Bachelor of Music Honours, Edith Cowan University]. Abgerufen von: [https://ro.ecu.edu.au/theses\\_hons/1530](https://ro.ecu.edu.au/theses_hons/1530)
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Networks. *arXiv:1406.2661*. Abgerufen von: <http://arxiv.org/abs/1406.2661>
- Greengard, S. (2019). Will deepfakes do deep damage? *Communications of the ACM*, 63(1), 17–19. <https://doi.org/10.1145/3371409>
- Guera, D., & Delp, E. J. (2018). Deepfake Video Detection Using Recurrent Neural Networks. *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 1–6. <https://doi.org/10.1109/AVSS.2018.8639163>



- Hassan, M. (2020, Oktober 29). *Deepfake Tutorial and Explanation Step by Step GPU/CPU (2020)*. Abgerufen von: <https://www.youtube.com/watch?v=tW7E-ENTWXRk>
- Hauser, A. (2021, März 18). *So funktionieren Audio-Deepfakes*. Abgerufen von: <https://www.scip.ch/?labs.20210318>
- Heath, N. (2020, Dezember 11). *What is AI? Everything you need to know about Artificial Intelligence*. Abgerufen von: <https://www.zdnet.com/article/what-is-ai-everything-you-need-to-know-about-artificial-intelligence/>
- Instagram. (2021, April 7). Abgerufen von: [https://www.instagram.com/felix\\_lobrecht/](https://www.instagram.com/felix_lobrecht/)
- iProov. (2020). *The Threat of Deepfakes: How biometrics protects consumers from the growing threat of deepfake attacks online*. iProov Limited. Abgerufen von: <https://www.iproov.com/blog/how-biometrics-protects-consumers-from-growing-threat-of-deepfakes>
- Jaiman, A. (2020, August 14). *Positive Use Cases of Deepfakes*. Abgerufen von: <https://towardsdatascience.com/positive-use-cases-of-deepfakes-49f510056387>
- Johnson, D. (2020, August 3). *Audio Deepfakes: Can Anyone Tell If They're Fake?* Abgerufen von: <https://www.howtogeek.com/682865/audio-deepfakes-can-anyone-tell-if-they-are-fake/>
- Juefei-Xu, F., Wang, R., Huang, Y., Guo, Q., Ma, L., & Liu, Y. (2021). *Countering Malicious DeepFakes: Survey, Battleground, and Horizon*. *arXiv:2103.00218*. Abgerufen von: <http://arxiv.org/abs/2103.00218>
- Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. *arXiv:1710.10196*. Abgerufen von: <http://arxiv.org/abs/1710.10196>

- Katarya, R., & Lal, A. (2020). A Study on Combating Emerging Threat of Deepfake Weaponization. *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 485–490. <https://doi.org/10.1109/I-SMAC49090.2020.9243588>
- Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat? *Business Horizons*, *63*(2), 135–146. <https://doi.org/10.1016/j.bushor.2019.11.006>
- Kim, B.-H., & Ganapathi, V. (2019). LumièreNet: Lecture Video Synthesis from Audio. *arXiv:1907.02253*. Abgerufen von: <http://arxiv.org/abs/1907.02253>
- Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Nießner, M., Pérez, P., Richardt, C., Zollhöfer, M., & Theobalt, C. (2018). Deep Video Portraits. *arXiv:1805.11714*. Abgerufen von: <http://arxiv.org/abs/1805.11714>
- Langa, J. (2021). Deepfakes, Real Consequences: Crafting Legislation To Combat Threats Posed by Deepfakes. *BOSTON UNIVERSITY LAW REVIEW*, *Vol. 101*(2), 761–801.
- Lee, D. (2019, Mai 10). *Deepfake Salvador Dalí takes selfies with museum visitors*. The Verge. Abgerufen von: <https://www.theverge.com/2019/5/10/18540953/salvador-dali-lives-deepfake-museum>
- Li, Y., Zhang, C., Sun, P., Qi, H., & Lyu, S. (2021). DeepFake-o-meter: An Open Platform for DeepFake Detection. *arXiv:2103.02018*. Abgerufen von: <http://arxiv.org/abs/2103.02018>
- Lorenzo-Trueba, J., Fang, F., Wang, X., Echizen, I., Yamagishi, J., & Kinnunen, T. (2018). Can we steal your vocal identity from the Internet?: Initial investigation of cloning Obama’s voice using GAN, WaveNet and low-quality found data. *arXiv:1803.00860*. Abgerufen von: <http://arxiv.org/abs/1803.00860>

- Lüscher, S. (2020, Februar 4). *Die Leiden der Schweizer Mediendemokratie*. SWI. Abgerufen von: [https://www.swissinfo.ch/ger/direktedemokratie/dunkelkammer-demokratie---serie\\_die-leiden-der-schweizer-mediendemokratie/45519640](https://www.swissinfo.ch/ger/direktedemokratie/dunkelkammer-demokratie---serie_die-leiden-der-schweizer-mediendemokratie/45519640)
- Mahmud, B., & Sharmin, A. (2020). Deep Insights of Deepfake Technology: A Review. *Dhaka University Journal of Science, Vol. 5(1)*, 13–23.
- Maksutov, A. A., Morozov, V. O., Lavrenov, A. A., & Smirnov, A. S. (2020). Methods of Deepfake Detection Based on Machine Learning. *2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, 408–411. <https://doi.org/10.1109/EIConRus49466.2020.9039057>
- Masood, M., Nawaz, M., Malik, K., Javed, A., & Irtaza, A. (2021). Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward. *arXiv:2103.00484*. Abgerufen von: <https://arxiv.org/abs/2103.00484>
- Melville, K. (2019, August 29). „*Humiliated, frightened and paranoid*“: *The insidious rise of deepfake porn*. Abgerufen von: <https://www.abc.net.au/news/2019-08-30/deepfake-revenge-porn-noelle-martin-story-of-image-based-abuse/11437774>
- Mirsky, Y., & Lee, W. (2021). The Creation and Detection of Deepfakes: A Survey. *ACM Computing Surveys, 54(1)*, 7:1-7:41. <https://doi.org/10.1145/3425780>
- Muna, M. (2020). Technological Arming: Is Deepfake the Next Digital Weapon? *UC Berkley*. Abgerufen von: [https://www.researchgate.net/publication/341781104\\_Technological\\_Arming\\_Is\\_Deepfake\\_the\\_Next\\_Digital\\_Weapon](https://www.researchgate.net/publication/341781104_Technological_Arming_Is_Deepfake_the_Next_Digital_Weapon)
- Nguyen, T. T., Nguyen, C. M., Nguyen, D. T., Nguyen, D. T., & Nahavandi, S. (2020). Deep Learning for Deepfakes Creation and Detection: A Survey. *arXiv:1909.11573*. Abgerufen von: <http://arxiv.org/abs/1909.11573>

- Pantserev, K. A. (2020). The Malicious Use of AI-Based Deepfake Technology as the New Threat to Psychological Security and Political Stability. In H. Jahankhani, S. Kendzierskyj, N. Chelvachandran, & J. Ibarra (Hrsg.), *Cyber Defence in the Age of AI, Smart Societies and Augmented Humanity* (S. 37–55). Springer International Publishing. [https://doi.org/10.1007/978-3-030-35746-7\\_3](https://doi.org/10.1007/978-3-030-35746-7_3)
- Perarnau, G., van de Weijer, J., Raducanu, B., & Álvarez, J. M. (2016). Invertible Conditional GANs for image editing. *arXiv:1611.06355*. Abgerufen von: <http://arxiv.org/abs/1611.06355>
- Perov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S., Umé, C., Dpfks, M., Facenheim, C. S., RP, L., Jiang, J., Zhang, S., Wu, P., Zhou, B., & Zhang, W. (2020). DeepFaceLab: A simple, flexible and extensible face swapping framework. *arXiv:2005.05535*. Abgerufen von: <http://arxiv.org/abs/2005.05535>
- Poulsen, K. (2019, Juni 1). We Found the Guy Behind the Viral ‘Drunk Pelosi’ Video. *The Daily Beast*. Abgerufen von: <https://www.thedailybeast.com/we-found-shawn-brooks-the-guy-behind-the-viral-drunk-pelosi-video>
- Project Revoice. (2021). Abgerufen von: <https://www.projectrevoice.org/>
- Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv:1511.06434*. Abgerufen von: <http://arxiv.org/abs/1511.06434>
- Reface. (2021). Abgerufen von: <http://hey.reface.ai/>
- Schick, N. (2020). *Deepfakes and the Infocalypse*. Octopus Publishing Group Ltd. Abgerufen von: [https://webreader.mytolino.com/library/index.html#/e-pub?id=DT0400.9781913183530\\_A40044115](https://webreader.mytolino.com/library/index.html#/e-pub?id=DT0400.9781913183530_A40044115)

- Sjouwerman, S. (2020). *The evolution of deepfakes: Fighting the next big threat*. Abgerufen von: <https://techbeacon.com/security/evolution-deepfakes-fighting-next-big-threat>
- Sohrawardi, S. J., Chintla, A., Thai, B., Seng, S., Hickerson, A., Ptucha, R., & Wright, M. (2019). Poster: Towards Robust Open-World Detection of Deepfakes. *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2613–2615. <https://doi.org/10.1145/3319535.3363269>
- Stucke, J. (2020, Dezember 3). Audio-Deepfakes—Was, wenn wir unseren Ohren nicht mehr trauen können? *Deutschlandfunk Kultur*. Abgerufen von: [https://www.deutschlandfunkkultur.de/audio-deepfakes-was-wenn-wir-unseren-ohren-nicht-mehr.976.de.html?dram:article\\_id=488600](https://www.deutschlandfunkkultur.de/audio-deepfakes-was-wenn-wir-unseren-ohren-nicht-mehr.976.de.html?dram:article_id=488600)
- Stupp, C. (2019, August 30). Fraudsters Used AI to Mimic CEO’s Voice in Unusual Cybercrime Case. *Wall Street Journal*. Abgerufen von: <https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402>
- Suwajanakorn, S., Seitz, S. M., & Kemelmacher-Shlizerman, I. (2017). Synthesizing Obama: Learning lip sync from audio. *ACM Transactions on Graphics*, 36(4), 1–13. <https://doi.org/10.1145/3072959.3073640>
- Tammekänd, J., Thomas, J., & Peterson, K. (2020). *Deepfakes 2020: The Tipping Point*. Abgerufen von: <https://thesentinel.ai/media/Deepfakes%202020:%20The%20Tipping%20Point,%20Sentinel.pdf>
- Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2016). Face2Face: Real-time Face Capture and Reenactment of RGB Videos. *arXiv:2007.14808*. Abgerufen von: <http://arxiv.org/abs/2007.14808>
- This Person Does Not Exist. (2021). Abgerufen von: <https://thispersondoesnotexist.com/>

- Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection. *arXiv:2001.00179*. Abgerufen von: <http://arxiv.org/abs/2001.00179>
- Tourismus Schweiz. (2021). *Sprachen in der Schweiz*. Abgerufen von: <https://www.myswitzerland.com/de-ch/planung/ueber-die-schweiz/zahlen-und-fakten/facts-about-switzerland/sprachen-in-der-schweiz/>
- Verhoeven, T. (2020). Die Risiken der Digitalisierung. In T. Verhoeven (Hrsg.), *Digitalisierung im Recruiting: Wie sich Recruiting durch künstliche Intelligenz, Algorithmen und Bots verändert* (S. 225–244). Springer Fachmedien. [https://doi.org/10.1007/978-3-658-25885-6\\_18](https://doi.org/10.1007/978-3-658-25885-6_18)
- WEMF AG. (2020). *WEMF Auflagebulletin 2020*. WEMF AG für Werbemedienforschung. Abgerufen von: [https://wemf.ch/media/wemf\\_aufgabebulletin\\_2020.pdf](https://wemf.ch/media/wemf_aufgabebulletin_2020.pdf)
- Westerlund, M. (2019). The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review*, 9, 39–52. <https://doi.org/10.22215/timreview/1282>
- Whitson, G. (2012, Januar 13). *What Are Magnet Links, and How Do I Use Them to Download Torrents?* Abgerufen von: <https://lifelifehacker.com/what-are-magnet-links-and-how-do-i-use-them-to-downloa-5875899>
- Whittaker, L., Kietzmann, T. C., Kietzmann, J., & Dabirian, A. (2020). “All Around Me Are Synthetic Faces”: The Mad World of AI-Generated Media. *IT Professional*, 22(5), 90–99. <https://doi.org/10.1109/MITP.2020.2985492>
- Wu, J., Feng, K., Chang, X., & Yang, T. (2020). A Forensic Method for DeepFake Image based on Face Recognition. *Proceedings of the 2020 4th High Performance Com-*

*puting and Cluster Technologies Conference & 2020 3rd International Conference on Big Data and Artificial Intelligence*, 104–108.  
<https://doi.org/10.1145/3409501.3409544>

Yang, Y., & Goh, B. (2019, November 29). China seeks to root out fake news and deepfakes with new online content rules. *Reuters*. Abgerufen von: <https://www.reuters.com/article/us-china-technology-idUSKBN1Y30VU>

Zhang, T., Deng, L., Zhang, L., & Dang, X. (2020). Deep Learning in Face Synthesis: A Survey on Deepfakes. *2020 IEEE 3rd International Conference on Computer and Communication Engineering Technology (CCET)*, 67–70.  
<https://doi.org/10.1109/CCET50901.2020.9213159>

Zucconi, A. (2018, März 14). *Understanding the Technology Behind DeepFakes*. Abgerufen von: <https://www.alanzucconi.com/2018/03/14/understanding-the-technology-behind-deepfakes/>

# Anhang

## A Deepfake-Anleitung der Community

### DeepFaceLab – Guide

Explanation of all DFL functions:

DeepFaceLab 2.0 consists of several .bat files used to perform various tasks/steps of creating a deepfake, they are located in the main folder along with two subfolders:

- `_internal` - internal files
- `workspace` - this is where your models, videos, datasets and final video outputs are

Terminology:

**Dataset (faceset)** - is a set of images that have been extracted (or aligned) from frames (extracted from video) or photos.

There are **two datasets** being used in **DFL 2.0** and they are **data\_dst** and **data\_src**:

- "**data\_dst**" is a folder that holds frames extracted from `data_dst.mp4` file - target video onto which we swap faces. It also contains 2 folders that are created after running face extraction. "aligned" containing images of faces (with embedded facial landmarks data) "aligned\_debug" which contains original frames with landmarks overlaid on faces which is used to identify correctly/incorrectly aligned faces (and it doesn't take a part in training or merging process). After cleaning up the dataset it can be deleted. Generates always for DST.

- "**data\_src**" is a folder that holds frames extracted from `data_src.mp4` file or where you can place images of your source - the person whose face you want to swap onto the target video. As with `data_dst`, after extracting faces 2 folders are created: "aligned" containing images of faces. "aligned\_debug" serves the same function as for DST however for SRC dataset extraction it is not generated by default, if you want it you need to select yes (y) when starting extraction to generate it. It can be also created after extraction.

Before you get to extract faces however you must have something to extract them from:



- **for data\_dst** you should prepare the target (destination) video and name it data\_dst.mp4  
- **for data\_src** you should either prepare the source video (as in examples above) and name it data\_src.mp4 or prepare images in jpg format.  
The process of extracting frames from video is also called extraction so for the rest of the guide/tutorial I'll be referring to both processes as "face extraction" and "frame extraction".

### 1. Workspace cleanup/deletion:

**1) Clear Workspace** - deletes all data from the "workspace" folder, feel free to delete this .bat file to prevent accidental removal of your workspace.

### 2. Frames extraction from source video (data\_src.mp4):

**2) Extract images from video data\_src** - extracts frames from data\_src.mp4 video and puts them into automatically created "data\_src" folder, available options:

- FPS - skip for videos default frame rate, enter numerical value for other frame rate (for example entering 5 will only render the video as it was 5 frames per second, meaning less frames will be extracted)

- JPG/PNG - choose the format of extracted frames, jpgs are smaller and generally have good enough quality so they are recommended, pngs are large and don't offer significantly higher quality but they are an option however because DFL uses JPGs only for training you should not use PNGs at all.

### 3. Video cutting (optional):

**3) cut video (drop video on me)** - allows to quickly cut any video to desired length by dropping it onto that .bat file. Useful if you don't have video editing software and want to quickly cut the video, options:

- From time - start of the video

- End time - end of the video

-Audio track - leave at default

Bitrate - let's you change bitrate (quality) of the video - best to be left at default

3. Frames extraction from destination video (data\_dst.mp4):

**3) extract images from video data\_dst FULL FPS** - extracts frames from data\_dst.mp4 file and puts them into newly created "data\_dst" folder, available options:

- JPG/PNG - same as in 2)

4. Data\_src faces extractin/alignment:

First stage of preparing source dataset is to extract faces from the extracted frames located inside "data\_src" folder.

There are 2 options:

**4) data\_src faceset extract MANUAL** - manual extractor, see 5.1 for usage.

**4) data\_src faceset extract** - automated extractor using S3FD algorithm - Use this first.

Available options for **S3FD** and **MANUAL** extractor are:

- Face Type:

a) full face (for FF models or lower: HF and MF, not recommended as it limits you to FF area of coverage)

b) whole face (for WF models or lower, recommended as an universal solution for working with both FF and WF models)

c) head (for HEAD models, not recommended for anything else, even WF)

- Resolution of the dataset, you can read more about it here: [REDACTED]

- Which GPU (or CPUT) to use for extraction (use GPU, it's almost always faster)

- Choosing whether to generate "aligned\_debug" images or not (can be generated afterwards)

After that is finished next step is to clean the source faceset/dataset of false positives/incorrectly aligned faces, for a detailed info check DATASET/FACASET part of the guide: [REDACTED]

**4.1) data\_src view aligned result** - opens up external app that allows to quickly go through the contents of "data\_src/aligned" folder for false positives and incorrectly aligned source faces as well as faces of other people so you can delete them.

**4.2) data\_src sort** - contains various sorting algorithms to help you find unwanted faces, these are the available options:

- [0] blur*
- [1] motion blur*
- [2] face yaw direction*
- [3] face pitch direction*
- [4] face rect size in source image*
- [5] histogram similarity*
- [6] histogram dissimilarity*
- [7] brightness*
- [8] hue*
- [9] amount of black pixels*
- [10] original filename*
- [11] one face in image*
- [12] absolute pixel difference*
- [13] best faces*
- [14] best faces faster*

**4.2) data\_src util add landmarks debug images** - let's you generate "aligned\_debug" folder after extracting faces (if you wanted to have it but forgot or didn't select the right option in the first place).

**4.2) data\_src util faceset enhance** - uses special machine learning algorithm to up-scale/enhance the look of faces in your dataset, useful if your dataset is a bit blurry or you want to make a sharp one have even more detail/texture.

**Optionally for enhancing SRC sets (not recommended for DST) you can use DFDNet**  
- Colab link here: 

**4.2) data\_src util faceset metadata restore** and **4.2) data\_src util faceset metadata save** - let's you save and restore embedded alignment data from your source faceset/dataset so you can edit some face images after you extracted them (for example sharpen them, edit out glasses, skin blemishes, color correct) without losing alignment data. EDITING ANY IMAGES FROM "ALIGNED" FOLDER WITHOUT THIS STEP WILL REMOVE THAT ALIGNMENT DATA AND THOSE PICTURES WON'T BE USABLE IN TRAINING, WHEN EDITING KEEP THE NAMES THE SAME, NO

FLIPPING/ROTATION IS ALLOWED, ONLY SIMPLE EDITS LIKE COLOR CORRECTION, OR RESIZING/UPSCALING ETC.

**4.2) data\_src util faceset pack** and **4.2) data\_src util faceset unpack** - packs/unpacks all faces from "aligned" folder into/from one file. Used for preparing custom pretraining dataset, easier sharing as one file and greatly improves dataset load times (seconds instead of minutes).

**4.2) data\_src util faceset resize** - allows you to resize your datasets to match resolution of your model, thus reducing CPU load during training and slightly improving performance.

**Make sure you backup your original dataset before resizing it as this process is irreversible!**

**4.2.other) data\_src util recover original filename** - reverts names of face images back to original order/filename (after sorting). Optional, training and merging will run correctly regardless of the SRC faces file names.

## 5. Data\_dst preparation:

Here steps are pretty much the same as with source dataset, with few exceptions, let's start with faces extraction/alignment process. We still have Manual and S3FD extraction method but there is also one that combines both and a special manual extraction mode, "aligned\_debug" folder is generated always.

**5) data\_dst faceset extract MANUAL RE-EXTRACT DELETED ALIGNED\_DEBUG** - manual re-extraction from frames deleted from "aligned\_debug" folder. More on that in 5. Data\_dst cleanup. Usage below in step 5.1.

**5) data\_dst faceset extract MANUAL** - manual extractor, see 5.1 for usage.

**5) data\_dst faceset extract + manual fix** - automated + manual extractor for frames where algorithm couldn't properly detect faces.

**5) data\_dst faceset extract** - automated extraction using S3FD algorithm.

Available options for all extractor modes are:

- choosing coverage area of extraction depending on face type of the model you want to train:

a) full face (for half, mid-half and full face)

b) whole face (for whole face but also works with others)

c) head (for head type of model)

- choosing which GPU (or CPU) to use for faces extraction/alignment process.

### 5.1 Manual extractor usage:

Upon starting the manual extractor or re-extractor a window will open up where you can manually locate faces you want to extract/re-extract:

- use your mouse to locate face

- use mouse wheel to change size of the search area

- make sure all or at least most landmarks (in some cases depending on the angle, lighting or present obstructions it might not be possible to precisely align all landmarks so just try to find a spot that covers all the visible bits the most and isn't too misaligned) land on important spots like eyes, mouth, nose, eyebrows and follow the face shape correctly, an up arrow shows you where is the "up" or "top" of the face

- use key A to change the precision mode, now landmarks won't "stick" so much to detected faces but you might be able to position landmarks more correctly

- user < and > keys (or , and .) to move back and forwards, to confirm a detection either left mouse click and move to the next one or hit enter

- right mouse button for aligning undetectable forward facing or non human faces (requires applying xseg for correct masking)

- q to skip remaining faces and quit extractor (it will also close down when you reach the last face and confirm it)

### 5.2 Data\_dst cleanup:

After we aligned data\_dst faces we have to clean them up, similar to how we did it with source faceset/dataset we have a selection of sorting methods which I'm not going to explain as they work exactly the same as ones for src. However cleaning up the destination dataset is different than source because we want to have all the faces aligned for all the frames where they are present - including obstructed ones. There are couple of tools at our disposal for that:

**5.1) data\_dst view aligned results** - let's you view the contents of "aligned" folder using external app (built into DFL) which offers quicker thumbnail generation than default windows explorer.

**5.1) data\_dst view aligned\_debug results** - let's you quickly browse contents of "aligned\_debug" folder to locate and delete any frames where our target person face has incorrectly aligned landmarks or where landmarks weren't placed at all (which means face wasn't detected at all). In general you use this to find if all your faces are properly extracted and aligned (if landmarks on some frames aren't lining up with the shape of the face or eyes/nose/mouth/eyebrows or are missing - they should be deleted so we can later manually re-extract/align them).

**5.2) data\_dst sort** - same as with source faceset/dataset, this tool let's you sort all aligned faces within "data\_dst/aligned" folder so that's it's easier to locate incorrectly aligned faces, false positives and faces of other people we don't want to train our model on/swap faces onto

**5.2) data\_dst util faceset pack** and **5.2) data\_dst util faceset unpack** - same as with source, let's you quickly pack entire dataset into one file.

**4.2) data\_src util faceset resize** - works the same as one for SRC dataset. **Remember to make a backup of your original dataset before resizing it as this process is irreversible!**

**5.2) data\_dst util recover original filename** - same as with source, restores original names/order of all aligned faces after sorting.

Start by sorting the dataset using **5.2) data\_dst sort** and use option 5: histogram similarity, this will allow you to find all the false positives (non faces, rotated faces, faces that are cropped too much/too little, faces of other people), to browse the set efficiently use **5.1) data\_dst view aligned results**, after you're done removing all bad faces revert the set back to original order using **5.2) data\_dst util recover original filename**.

Next move on to **5.1) data\_dst view aligned\_debug results** where you will be able to browse through all of the aligned\_debug frames, these don't take part in training so after sorting they can be remove, before you do that however they're essential to re-extract all the missing faces and those we've just deleted, browse through all of them and select all frames where landmarks don't exist, are placed in the wrong place or are placed over the face we're extracting but not in a correct way (misalignment of landmarks), after selecting them all press delete to remove them and now you can close XNView app.

Last step of clean up is to manually re-extract all of those missing faces, to do so use **5) data\_dst faceset extract MANUAL RE-EXTRACT DELETED ALIGNED\_DEBUG** which will check all deleted debug frames and open up manual alignment tool where you will have to use your mouse to manually locate the face, adjust the size of detection square until the landmarks snap onto facial features and outline everything precisely, including the outline (shape) of the jaw.

To adjust size of the detection square use mouse wheel, you can move the square simply by moving the mouse around the screen, if you can't find good spot that aligns everything correctly press "A" to change precision mode and try to find good alignment in this mode, if this one doesn't work switch back to default mode. To confirm selection you can either press enter, or do so manually by clicking with left mouse button to confirm the landmarks, you can use <> arrows (, . on standard qwerty keyboard layout) to move back and forward. Align manually all faces and once you reach the last image and hit enter or next frame the window will close and on command line windows you will see the progress of faces being saved, once the process finishes simply click on the command line windows and press enter to close it. Now you should have all faces extracted but in some cases you will have to run it few times (for example if your faces as a mirror reflection, or in case of some kind of transition where temporarily both faces are visible). I will not go into more detail on how to extract multiple people, simply separate both faces into separate folders and perform re-extraction on each set separately, you'll do it twice or two faces, 3 times for 3, and so on.

### 5.3 XSEG model training and faceset marking:

**NEW:** There is now a pretrained Generic WF XSeg model included with DFL, if you don't want to label your own faces and train XSeg model yourself you can use it. This is not for full face so if that's face type you're using you will still have to label your own faces and train the model, same applies to head models, they too still require labeling and training of your own XSeg model.

**5.XSeg Generic) data\_dst whole\_face mask - apply** - applies WF masks to your DST dataset.

**5.XSeg Generic) data\_src whole\_face mask - apply** - applies WF masks to your SRC dataset.

XSeg is a replacement for defunct FANSeg model that is used to automatically mask your result face on top of your target video. It is also used to make obstructions such as hair or hands visible over the swapped/fake face and is completely customizable by the means of manual dataset marking and model training. Such models can be also reused similarly to the swapping models (SAEHD, Quick96) so when you start working on a new video you don't need to train the model from scratch but instead can reuse existing one by feeding it marked (labeled) faces. Both SRC and DST datasets can be marked thus giving you options of using XSeg-prd (respects SRC face shape) and XSeg-dst (respects DST face shape) masking modes and also combine them both. It is also possible to combine both of them with learned mask (see the merger masking modes, merging step). XSeg works with all face types such as Full Face, Whole Face and even Head so you have full control of which parts of the DST faces get covered/swapped with new face and which parts (obstructions) are being left visible.

**5.XSeg) data\_dst mask for XSeg trainer - edit** - label tool to mark destination faces with XSeg polygons.

**5.XSeg) data\_dst mask for XSeg trainer - fetch** - copies faces containing XSeg polygons to folder "aligned\_xseg". Can be used to collect labeled faces so they can be reused in future XSeg model trainings.

**5.XSeg) data\_dst mask for XSeg trainer - remove** - removes labeled/marked XSeg polygons from the extracted frames.

**5.XSeg) data\_src mask for XSeg trainer - edit** - label tool to mark source faces with XSeg polygons.

**5.XSeg) data\_src mask for XSeg trainer - fetch** - copies faces containing XSeg polygons to folder "aligned\_xseg". Can be used to collect labeled faces so they can be reused in future XSeg model trainings.

**5.XSeg) data\_src mask for XSeg trainer - remove** - removes labeled/marked XSeg polygons from the extracted frames.

**XSeg) train.bat** - runs the training of the XSeg model.



**5.XSeg.optional) trained mask for data\_dst - apply** - replaces default DST masks derived from landmarks created during extraction with ones generated by the trained XSeg model, it is required for proper whole face and head face type model training and also if you plan on using style power with those 2 face types.

**5.XSeg.optional) trained mask for data\_dst - remove** - removes XSeg masks and restores default DST masks.

**5.XSeg.optional) trained mask for data\_src - apply** - replaces default DST masks derived from landmarks created during extraction with ones generated by the trained XSeg model, it is required for proper whole face and head face type model training and also if you plan on using style power with those 2 face types.

**5.XSeg.optional) trained mask for data\_src - remove** - removes XSeg masks and restores default DST masks. Before you start it's important to know the difference between face marking/labeling and masking. Marks/labels are polygons you create manually in the editor which model uses to learn how to mask faces, masks are what gets applied by the apply .bat and also what merger will generate using your trained XSeg model during merging. Masks define which area on the face sample is the face itself and what is a background or obstruction, that's why you need to apply them for WF and HEAD model as default masks don't cover the area needed for those face types and also why even with full face I recommend to apply XSeg to both SRC and DST as XSeg masks are much more precise and true to face shape than extraction generated DST masks and even learned masks that model trains based on landmarks of SRC and DST faces to create learned-prd and learned-dst masks respectively during training (they don't understand face obstructions and if your landmarks are off/incorrect - so will be those learned masks).

## **XSeg usage:**

### **1. Mark your datasets**

Start by marking both SRC\* and DST faces using 5.XSeg) data\_src mask for XSeg trainer - edit and 5.XSeg) data\_dst mask for XSeg trainer – edit

*Each tool has a written description that's displayed when you go over it with your mouse (en/ru/zn languages are supported).*

Mark 50 to 200 different faces for both SRC and DST, you don't need to mark all faces but only those where the face looks significantly different, for example:  
- when facial expression changes (open mouth - closed mouth, big smile - frown)

- when direction/angle of the face changes
- or when lighting conditions/direction changes (usually together with face angle but in some cases the lighting might change while face still looks in the same direction). The more various faces you mark, the better quality masks Xseg model will generate for you. In general the smaller the dataset is the less faces will have to be marked and the same goes about the variety of angles, if you have many different angles and also expressions it will require you to mark more faces.

Keep the same "logic" of marking for all faces, for example:

- the same approximated jaw line of the side faces, where the jaw is not visible
- the same hair line

Marking obstructions:

While marking faces you will also probably want to exclude obstructions so that they are visible in the final video, to do so you can either:

- not include obstructions in the main polygon that defines face area you want to be swapped.
- or use exclude poly mode to draw additional label around the obstruction or part you want to be visible/not swapped.

When marking obstructions you need to make sure you label them on several faces according to the same rules as when marking faces with no obstructions, mark the obstruction (even if it doesn't change appearance/shape/position when face/head:

- changes angle
- facial expression changes
- lighting conditions change

*If the obstruction is additionally changing shape and/or moving across the face you need to mark it few times, not all obstructions on every face need to be labeled though but still the more variety of different obstructions occur in various conditions - the more faces you will have to label.*

How to mark faces for different face types:

- For FF mark faces from chin up to slightly above eyebrows, if you're using FF faces then on some chin will be cut off, in that case you can round the edge of as to not have

perfectly straight lines on some parts of the face, remember as with SAEHD model training you can use higher face type faces and train lower face type model, you can even use WF faces, train WF XSeg model but mark them in a way FF face would be marked. Profiles are almost always not fully covered with FF face type so make sure to also not have perfectly straight lines of your label on the parts that get cut off.

- For WF mark faces from chin up to the hairline for frontal shots, for profile shots make sure to follow jawline (if it's pronounced it will be visible, if it's not as visible approximate it and use the same logic for marking of all faces), don't include ears if they are visible, mark up to where hair starts.

- For HEAD include the whole face as well as hair, make sure the masks are precise, include ears, optionally you can also include a little bit of the neck.

Use exclude tool on all major and some minor obstructions:

- large/thick hair strands and some smaller individual hairs if they're thick/noticeable enough (lots of contrast against skin)

- hands and other objects held in front of the face (for NSFW obviously mark out all other obvious objects)

- shoulders, legs and other large obstructions that cover significant part of the face (make sure landmarks for this face are correct on the visible parts and also fairly accurate in position and shape on the obstructed parts)

- tongues when they are sticking out of the mouth

- for DST faces exclude mouth cavity when they are wide open (including tongue inside), if teeth aren't visible mark it to where the lips start, if teeth are visible make sure to leave a decent offset to prevent double teeth after blurring the mask in the merger, for SRC faces you can skip this step and include entire face, including tongue, mouth cavity and all imperfections you want trained and carried over to the final predicted face (but you may exclude piercing and other things on the face that are temporary/only exist on some faces, don't exclude beauty marks and other parts of the face that are characteristic of this person)

- for NSFW faces, mark out/exclude any liquids, if the liquid looks like water and isn't thick enough (or doesn't contrast with the face enough or is very small (pixel sized) do not exclude it, it will confuse the model during training and it will treat all light reflections/skin highlights as liquids which may end up showing too much of the DST face, especially if you blur the mask heavily during merging and there will be bright spots in crucial areas like mouth, nose, eyes or teeth.

Once you finish marking DST faces scroll to the end of the list and hit Esc to save them, then you can move on to training your XSEG model.

\*If you're training a full face model you may skip marking of SRC faces and only do DST faces however if you plan on using Style Powers it's recommended to also mark SRC faces and as I've already explained you need to apply XSeg masks to both SRC and DST datasets before training (or at least before using Style Powers). Also if you want to use XSeg-prd you must mark your SRC faces, otherwise the model won't be able to correctly create masks that respect the shape of SRC faces.

## **2. Train your XSeg model.**

When starting training for the first time you will see an option to select face\_type of the XSeg model which should be set to be the same as face\_type of your face swapping model although you can also use a higher option safely. Reason why you can use higher option is that even if you have a whole face dataset, marked as whole face set should be with which you train a WF XSeg and then you use a full face face\_type SAEHD model, while merging the mask will be cropped to the highest area possible by the full face face\_type. If you then take the same dataset but train an FF XSeg model the mask will be still cropped to the same size and they will be virtually the same, that's why I recommend to stick with WF XSeg for both FF and WF SAEHD models.

You will also be able to choose device to train on as well as batch size which will typically be much higher as XSeg model is not that demanding as training of the face swapping model (you can also start off at lower value and raise it later).

You can switch preview modes using space (there are 3 modes, DST training, SRC training and SRC+DST (distorted)).

To update preview progress press P.

Esc to save and stop training.

During training check previews often, if some faces have bad masks after about 50k iterations (bad shape, holes, blurry), save and stop training, apply masks to your dataset, run editor, find faces with bad masks by enabling XSeg mask overlay in the editor, mark them and hit esc to save and exit and then resume XSeg model training, when starting up an already trained model you will get a prompt if you want to restart training, select no (n) as selecting yes (y) will restart the model training from 0 instead of continuing. However in case your masks are not improving despite having marked many more faces and being well above 100k-150k iterations it might be necessary to mark even more faces or restart training from scratch.

### **3. Apply XSeg masks to your datasets.**

This step has been already explained few times here but in any case I'm repeating it so it's all clear. After you're done training or after you've already applied XSeg once and then fixed faces that had bad masks it's time for final application of XSeg masks to your datasets. Also as I've already explained it is not necessary to apply datasets if you're using a full face SAEHD model, you can simply use XSeg during merging by selecting the new masking modes such as XSeg-Prd or XSeg-Dst (or combinations of them with themselves and learned masks) but for best results it's recommended to apply them anyway as you will get better results when using Style Power because those XSeg masks are much more precise than learned masks that are otherwise used to define areas of face and background for Style Transfer training.

#### **Extra tips:**

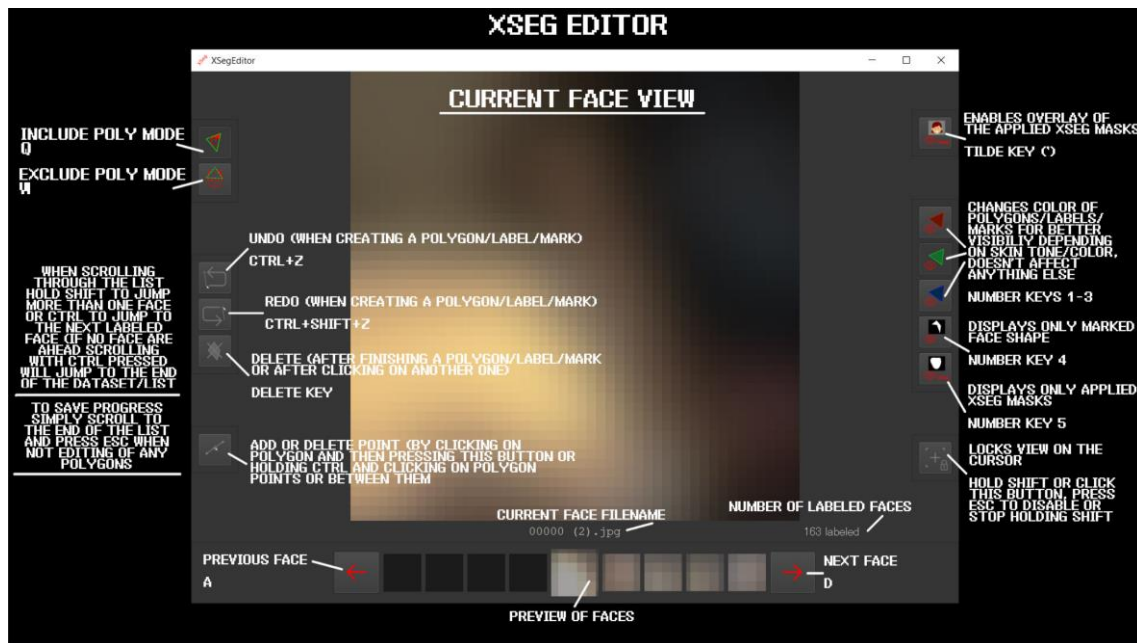
1. Don't bother making 1000 point label, it will take too much time to mark all the faces and won't affect the face vs if you use just 30-40 points to describe the face shape but also don't try to mark it with 10 points or the mask will not be smooth, the exception here would be marking hair for HEAD face type training where obviously some detail is needed to correctly resolve individual hair strands.
2. Do not mark shadows unless they're pitch black.
3. Don't mark out tongues or insides of the mouth if it's barely open.
4. If obstruction or face is blurry mark as much as needed to cover everything that should or shouldn't be visible, do not make offsets too big.
5. Keep in mind that when you use blur the edge blurs both in and out, if you mark out a finger right on the edge it won't look bad on low blur but on higher one it will start to

disappear and be replaced with the blurry version of what model learned, same goes for the mouth cavity, on low blur it will only show result face teeth but if you apply high blur then DST teeth will start to show and it will look bad (double teeth).

This means:

- when excluding obstructions like fingers - mark it on the edge or do a tiny offset. Both SRC and DST
- when excluding mouth cavity - mark it with bigger offset from teeth. DST, SRC is optional
- when excluding tongues - mark it on the edge, offset from teeth, only when it sticks out significantly. Both SRC and DST

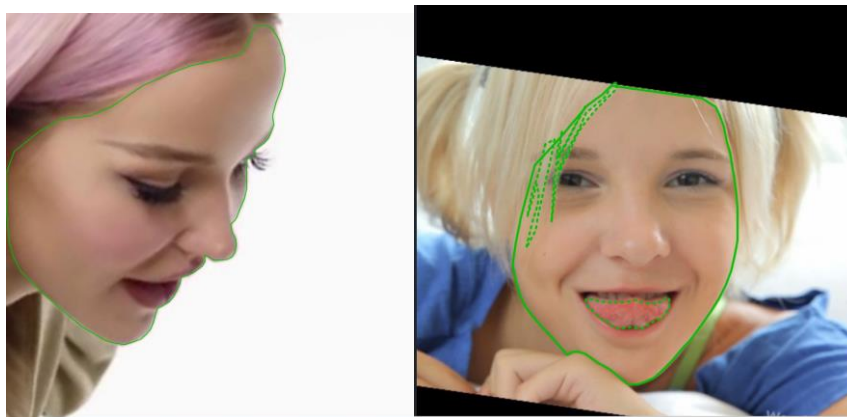
### XSeg editor:



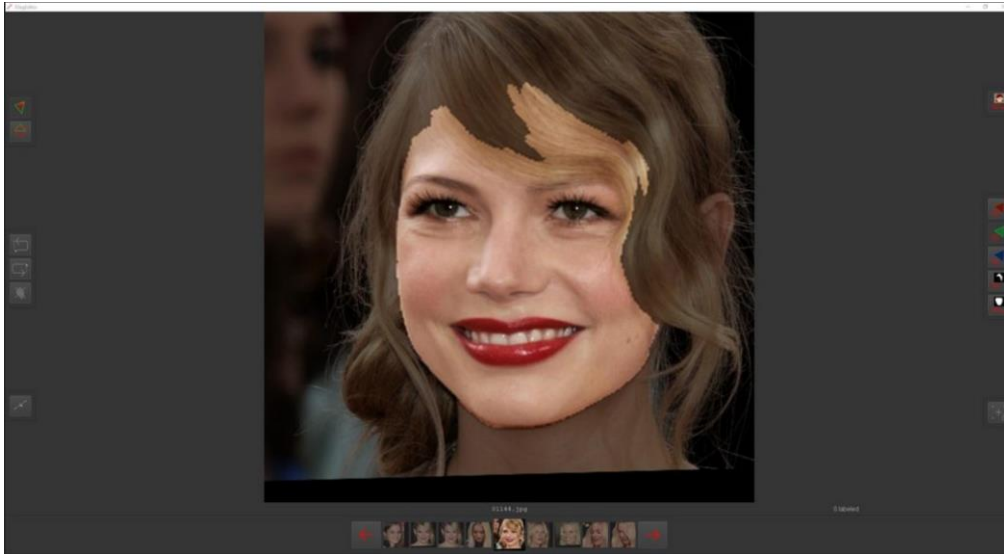
## Training preview:



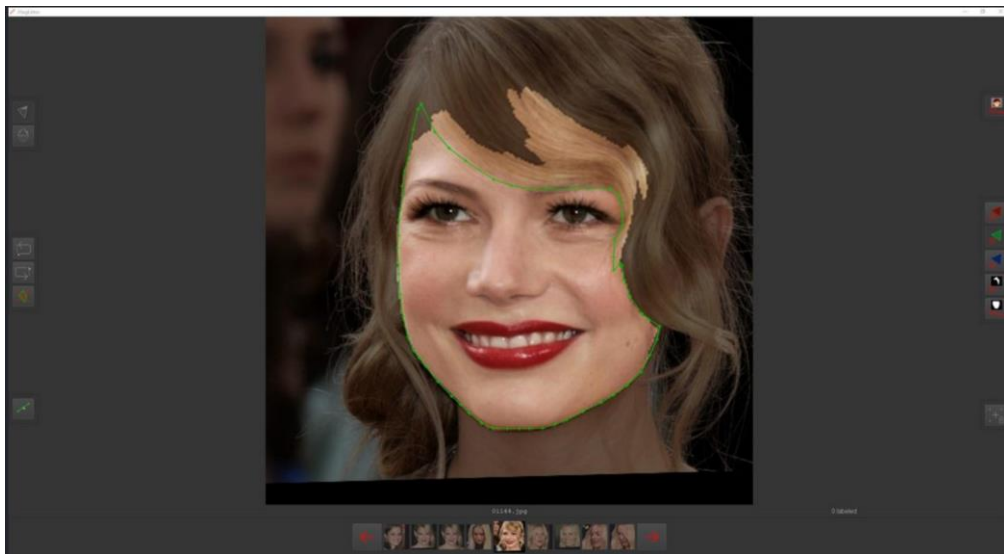
## Some correctly masked faces examples:



**Example of faces with bad applied mask:**



**Fixing the mask by marking those faces correctly:**



How to use shared marked faces to train your own XSeg model:

Download, extract and place faces into "data\_src/aligned" or "data\_dst/aligned". Make sure to rename them to not overwrite your own faces (I suggest XSEGSRC and XSEG DST for easy removal afterwards). You can mix shared faces with your own labeled to give the model as much data to learn masks as possible, don't mix face types, make sure all faces roughly follow the same logic of masking. Then just start training your XSeg model (or shared one).



How to use shared XSeg model and apply it to your dataset:

Simply place it into the "model" folder and use apply .bat files to apply masks to SRC or DST. After you apply masks open up XSeg editor and check how masks look by enabling XSeg mask overlay view, if some faces don't have good looking masks, mark them, exit the editor and start the training of the XSeg model again to fix them. You can also mix in some of the shared faces as described above (how to use shared marked faces). You can reuse XSeg models (like SAEHD models).

User shared SAEHD models can be found in this thread: [\[REDACTED\]](#)

## 6. Training:

**There are currently 2 models to choose from for training:**

**SAEHD (6GB+):** High Definition Styled Auto Encoder - for high end GPUs with at least 6GB of VRAM. Adjustable

**Quick96 (2-4GB):** Simple mode dedicated for low end GPUs with 2-4GB of VRAM.

Fixed parameters:

- 96x96 Pixels resolution
- Full Face
- Batch size 4
- DF-UD architecture

Quick96 is recommended for low end cards or dataset testing only. Model settings spreadsheet [\[REDACTED\]](#). After you've checked other peoples settings and decided on a model you want to use you start it up using one of those:

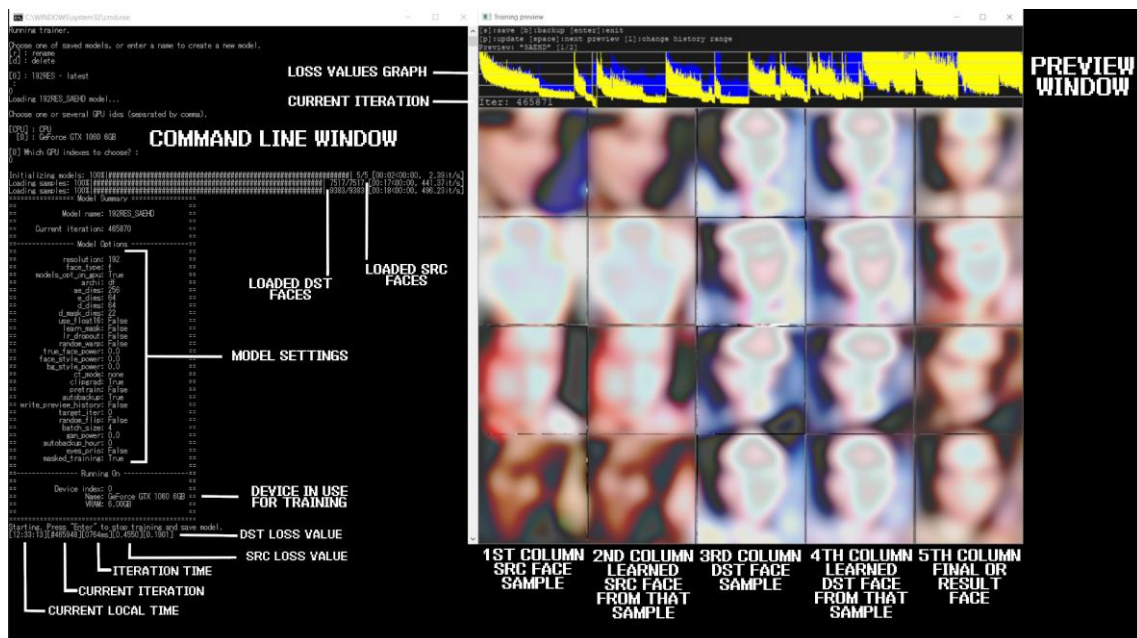
**6) train SAEHD**

**6) train Quick96**

Since Quick96 is not adjustable you will see the command window pop up and ask only 1 question - CPU or GPU (if you have more then it will let you choose either one of them or train with both). SAEHD however will present you with more options to adjust.

In both cases first a command line window will appear where you input your model settings. On a first start will you will have access to all setting that are explained below, on startup of training with a model already trained and present in the "model" folder you will also receive a prompt where you can choose which model to train on (if you have more than one set of model files present in your "model" folder). You will also always get a prompt to select which GPU or CPU you want to run the trainer on.

Second thing you will see once you startup is the preview window that looks like this:



Here is a more detailed explanation of all functions in order they are presented to the user upon starting training of a new model:

*Note that some of these get locked and can't be changed once you start training due to way these models work, example of things that can't be changed later are:*

- model resolution
- model architecture
- models dimensions (dims)
- face type


**Autobackup every N hour ( 0..24 ?:help )** : self explanatory - let's you enable automatic backups of your model every N hours. Leaving it at 0 (default) will disable auto backups. Default value is 0 (disabled).

**[n] Write preview history ( y/n ?:help )** : save preview images during training every few minutes, if you select yes you'll get another prompt: **[n] Choose image for the preview history ( y/n )** : if you select N the model will pick faces for the previews randomly, otherwise selecting Y will open up a new window after datasets are loaded where you'll be able to choose them manually.

**Target iteration** : will stop training after certain amount of iterations is reached, for example if you want to train you model to only 100.000 iterations you should enter a value of 100000. Leaving it at 0 will make it run until you stop it manually. Default value is 0 (disabled).

**[n] Flip SRC faces randomly ( y/n ?:help )** : Randomly flips SRC faces horizontally, helps to cover all angles present in DST dataset with SRC faces as a result of flipping them which can be helpful sometimes (especially if our set doesn't have many different lighting conditons but has most angles) however in many cases it will make results seem unnatural becasue faces are never perfectly symmetric, it will also copy facial features from one side of the face to the other one, they may then appear on either sides or on both at the same time. Recommended to only use early in the training or not at all if our SRC set is diverse enough. Default value is N.

**[y] Flip DST faces randomly ( y/n ?:help )** : Randomly flips DST faces horizontally, can improve generalization when **Flip SRC faces randomly** is disabled. Default value is Y.

**Batch\_size ( ?:help )** : Batch size settings affects how many faces are being compared to each other every each iteration. Lowest value is 2 and you can go as high as your GPU will allow which is affected by VRAM. The higher your models resolution, dimensions and the more features you enable the more VRAM will be needed so lower batch size might be required. It's recommended to not use value below 4. Higher batch size will provide better quality at the cost of slower training (higher iteration time). For the initial stage it can be set lower value to speed up initial training and then raised higher. Optimal values are between 6-12. How to guess what batch size to use? You can either use trial and error or help yourself by taking a look at what other people can achieve on their GPUs by checking out the DFL 2.0 spreadsheet: 



**Resolution ( 64-640 ?:help )** : here you set your models resolution, bear in mind this option cannot be changed during training. It affects the resolution of swapped faces, the higher model resolution - the more detailed the learned face will be but also training will be much heavier and longer. Resolution can be increased from 64x64 to 640x640 by increments of:

16 (for regular and -U architectures variants)

32 (for -D and -UD architectures variants)

Higher resolutions might require increasing of the model dimensions (dims).

**Face type ( h/mf/f/wf/head ?:help )** : this option let's you set the area of the face you want to train, there are 5 options - half face, mid-half face, full face, whole face and head:

**a) Half face (HF)** - only trains from mouth to eyebrows but can in some cases cut off the top or bottom of the face (eyebrows, chin, bit of mouth).

**b) Mid-half face (MHF)** - aims to fix HF issue by covering 30% larger portion of face compared to half face which should prevent most of the undesirable cut offs from occurring but they can still happen.

**c) Full face (FF)** - covers most of the face area, excluding forehead, can sometimes cut off a little bit of chin but this happens very rarely (only when subject opens mouth wide open) - most recommended when SRC and/or DST have hair covering forehead.

**d) Whole face (WF)** - expands that area even more to cover pretty much the whole face, including forehead and all of the face from the side (up to ears, HF, MHF and FF don't cover that much).

**e) Head (HEAD)** - is used to do a swap of the entire head, not suitable for subjects with long hair, works best if the source faceset/dataset comes from single source and both SRC and DST have short hair or one that doesn't change shape depending on the angle.

### **Examples of faces, front and side view when using all face types:**

**AE architecture (df/liae/df-u/liae-u/df-d/liae-d/df-ud/liae-ud ? :help )** : This option let's you choose between 2 main learning architectures: DF and LIAE as well as their -U, -D and -UD variants.

**DF:** This model architecture provides a more direct face swap, doesn't morph faces but requires that the source and target/destination face/head have similar shape while facial features (shapes of mouth, eyes, nose) can differ more than with LIAE. Works better on frontal shots and requires that your source dataset has all the required angles, can produce worse results on side profiles than LIAE.

**LIAE:** This model architecture isn't as strict when it comes to face/head shape similarity between source and target/destination but facial features (shapes of eyes, nose, mouth) should be similar between source and target/destination for good results. This model offers worse resemblance to source than DF but can handle side profiles better and is more forgiving when it comes to source faceset/dataset missing some angles, expressions or lighting conditions, often produces more refined face swaps with better color/lighting match.

**DF-U/LIAE-U:** this variant aims to improve similarity/likeness to the source faces.

**DF-D/LIAE-D:** this variant aims to improve performance by roughly doubling possible resolution with no extra compute cost (VRAM usage) and similar performance. However it requires longer training, model must be pretrained first for optimal results and resolution must be changed by the value of 32 as opposed to 16 in other variants.

**DF-UD/LIAE-UD:** combines both variants for maximum likeness and increased resolution/performance. Also requires longer training and model to be pretrained.

**The next 4 options control models neural network dimensions which affect models ability to learn, modifying these can have big impact on performance and quality:**

**AutoEncoder dims ( 32-1024 ?:help ) :** Auto encoder dims setting, affects overall ability of the model to learn faces.

**Encoder dims ( 16-256 ?:help ) :** Encoder dims setting, affects ability of the model to learn general structure of the faces.

**Decoder dims ( 16-256 ?:help ) :** Decoder dims setting, affects ability of the model to learn fine detail.

**Decoder mask dims ( 16-256 ?:help ) :** Mask decoder dims setting, affects quality of the learned masks. May or may not affect some other aspects of training.

The changes in performance when changing each setting can have varying effects on performance and it's not possible to measure effect of each one on performance and quality without extensive testing. Each one is set at certain default value that should offer optimal results and good compromise between training speed and quality. Also when changing one parameter the other ones should be changed as well to keep the relations between them similar, that means raising AE to 320 should be done while also raising E and D to 72 and D Mask can be raised but it's optional and can be left at default or lowered to 16. It's always best to raise them all when you're training higher model resolutions because one can't just raise resolution and expect significant improvement from just that.

**Masked training ( y/n ?:help ) :** Prioritizes training of what's masked (default mask or applied xseg mask), available only for WF and HEAD face types, disabling it trains the whole sample area (including background) at the same priority as the face itself. Default value is y (enabled).

**Eyes and mouth priority ( y/n ?:help ) :** Attempts to fix problems with eyes and mouth (including teeth) by training them at higher priority, can improve their sharpness/level of detail too.

**Uniform\_yaw ( y/n ?:help )** : Helps with training of profile faces, forces model to train evenly on all faces depending on their yaw and prioritizes profile faces, may cause frontal faces to train slower, enabled by default during pretraining, can be used while RW is enabled to improve generalization of profile/side faces or when RW is disabled to improve quality and sharpness/detail of those faces. Useful when your source dataset doesn't have many profile shots. Can help lower loss values. Default value is n (disabled).

**Place models and optimizer on GPU ( y/n ?:help )** : Enabling GPU optimizer puts all the load on your GPU which greatly improves performance (iteration time) but will lead to higher VRAM usage, disabling this feature will off load some work of the optimizer to CPU which decreases load on GPU and VRAM usage thus letting you achieve higher batch size or run more demanding models at the cost of longer iteration times. If you get OOM (out of memory) error and you don't want to lower your batch size or disable some feature you should disable this feature and thus some work will be offloaded to your CPU and you will be able to run your model without OOM errors at the cost of lower speed. Default value is y (enabled).

**Use AdaBelief optimizer? ( y/n ?:help )** : AdaBelief (AB) is a new model optimizer which increases model accuracy and quality of trained faces, when this option is enabled it replaces the default RMSProp optimizer. However those improvements come at a cost of higher VRAM usage, requiring existing or new models to be trained at lower batch size to prevent OOM errors. This option should be only used on new models and always from the start, never switch it off during training, once it's enabled it should remain so. When using AdaBelief LRD is optional but still recommended and should be enabled (LRD) before and while running GAN training. Default value is Y.

**Use learning rate dropout ( y/n/cpu ?:help )** : LRD is used to accelerate training of faces and reduces sub-pixel shake (reduces face shaking and to some degree can reduce lighting flicker as well). It's primarily used in 3 cases:

- before disabling RW, when loss values aren't improving by a lot anymore, this can help model to generalize faces a bit more

- after RW has been disabled and you've trained the model well enough enabling it near the end of training will result in more detailed, stable faces that are less prone to flicker
- after you've trained a bit with LRD and faces look as good as they can you enable GAN and LRD should be kept enabled while it runs.

This option affects VRAM usage so if you run into OOM errors you can run it on CPU at the cost of 20% slower iteration times or just lower your batch size.

For more detailed explanation of LRD and order of enabling main features during training please refer to **FAQ Question 8**. If your model has Adabelief enabled LRD is optional but still recommended, especially when training GAN.

**Enable random warp of samples ( y/n ?:help )** : Random warp is used to generalize a model so that it correctly learns face features and expressions in the initial training stage but as long as it's enabled the model may have trouble learning the fine detail - because of it it's recommended to keep this feature enabled as long as your faces are still improving (by looking at decreasing loss values and faces in the preview window improving) and once all look correct (and loss isn't decreasing anymore) you should disable it to start learning details, from then you don't re-enable it unless you ruin the results by applying to high values for certain settings (style power, true face, etc) or when you want to reuse that model for training of new target video with the same source or when reusing with combination of both new SRC and DST, you always start training with RW enabled. Default value is y (enabled).

**GAN power ( 0.0 .. 10.0 ?:help )** : GAN stands for Generative Adversarial Network and in case of DFL 2.0 it is implemented as an additional way of training to get more detailed/sharp faces. This option is adjustable on a scale from 0.0 to 10.0 and it should only be enabled once the model is more or less fully trained (after you've disabled **random warp of samples and enabled LRD**). It's recommended to use low values like 0.01. Make sure to backup your model before you start training (in case you don't like results, get artifacts or your model collapses). Once enabled two more settings will be presented to adjust internal parameters of GAN:

**[ ] GAN patch size ( 3-640 ?:help )** : Improves quality of GAN training at the cost of higher VRAM usage, default value is 1/8th of your resolution.



**[16] GAN dimensions ( 4-64 ?:help ) :** The dimensions of the GAN network. The higher the dimensions, the more VRAM is required but it can also improve quality, you can get sharp edges even at the lowest setting and because of this default value of 16 is recommended but you can reduce it to 12-14 to save some performance if you need to.

Before/after example of a face trained with GAN at value of 0.1 for 40k iterations:



**'True face' power. ( 0.0000 .. 1.0 ?:help ) :** True face training with a variable power settings let's you set the model discriminator to higher or lower value, what this does is it tries to make the final face look more like src, as with **GAN** this feature should only be enabled once random warp is disabled and model is fairly well trained. Consider making a backup before enabling this feature. Never use high values, typical value is 0.01 but you can use even lower ones like 0.001. The higher the setting the more result face will look like faces in source dataset which may cause issues with color match and also cause artifacts to show up so it's important to not use high values. It has a small performance impact which may cause OOM error to occur. Default value is 0.0 (disabled).



**Face style power ( 0.0..100.0 ? :help )** and **Background style power ( 0.0..100.0 ? :help )** : This setting controls style transfer of either face or background part of the image, it is used to transfer the style of your target/destination faces (data\_dst) over to the final learned face which can improve quality and look of the final result after merging but high values can cause learned face to look more like data\_dst than data\_src. It will transfer some color/lighting information from DST to result face which may help with color matching and reduce flickering if color transfer isn't enough. It's recommended to not use values higher than 10. Start with small values like 0.001-0.01 and increase it or start with higher value like 1-2 and decrease it over time. This feature has performance impact and using it will increase iteration time and may require you to lower your **batch size**, **disable gpu optimizer** or **run LRD on CPU** as a result of higher VRAM usage. Consider making a backup before enabling this feature. Default value is 0.0 (disabled).

**Color transfer for src faceset ( none/rct/lct/mkl/idt/sot ? :help )** : this features is used to match the colors of your data\_src to the data\_dst so that the final result has similar skin color/tone to the data\_dst and the final result after training doesn't change colors when face moves around (which may happen if various face angles were taken from various sources that contained different light conditions or were color graded differently, commonly referred to as flickering or flicker). There are several options to choose from:

- **none**: because sometimes less is better and in some cases you might get better results without any color transfer during training.
- **rct** (reinhard color transfer): based on: <https://www.cs.tau.ac.il/~turkel/imagepa...ansfer.pdf>

- **lct** (linear color transfer): Matches the color distribution of the target image to that of the source image using a linear transform.
- **mkl** (Monge-Kantorovitch linear): based on: <http://www.mee.tcd.ie/~sigmedia/pmwiki/u...tie07b.pdf>
- **idt** (Iterative Distribution Transfer): based on: <http://citeseerx.ist.psu.edu/viewdoc/dow...1&type=pdf>
- **sot** (sliced optimal transfer): based on: <https://dcoeurjo.github.io/OTColorTransfer/>

**Enable gradient clipping ( y/n ?:help )** : This feature is implemented to prevent so called model collapse/corruption which may occur when using various features of DFL 2.0. It has small performance impact so if you really don't want to use it you must enable auto backups as a collapsed model cannot recover and must be scraped and training must be started all over. Default value is n (disabled) but since the performance impact is so low and it can save you a lot of time by preventing model collapse if you leave it enabled. Model collapse is most likely to happen when using Style Powers so if you're using them it's highly advised to enable gradient clipping or backups (you can also do them manually).

**Enable pretraining mode ( y/n ?:help )** : Enables pretraining process that uses a dataset of random peoples faces to initially train your model, after training it to around 200k-400k iterations such model can be then used when starting training with actual data\_src and data\_dst you want to train, it saves time because you don't have to start training all over from 0 every time (the model will "know" how faces should look like and thus speed up the initial training stage).

## 7. Merging:

**After you're done training your model it's time to merge learned face over original frames to form final video (convert).**For that we have 2 converters corresponding to 2 available models:

## 7) merge SAEHD

## 7) merge Quick96

Upon selecting any of those a command line window will appear with several prompts. 1st one will ask you if you want to use an interactive converter, default value is y (enabled) and it's recommended to use it over the regular one because it has all the features and also an interactive preview where you see the effects of all changes you make when changing various options and enabling/disabling various features. **Use interactive merger? ( y/n ) :**

2nd one will ask you which model you want to use: **Choose one of saved models or enter a name to create a new model.**

*[r] : rename*

*[d] : delete*

*[0] : df192 – latest*

3rd one will ask you which GPU/GPUs or CPU you want to use for the merging (conversion) process: **Choose one or several GPU idxs (separated by comma).**

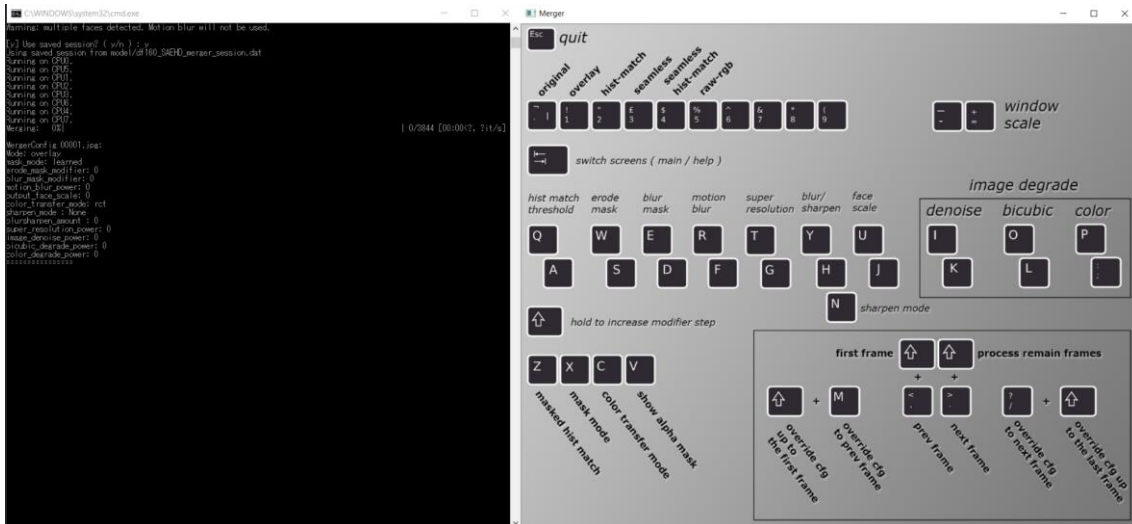
*[CPU] : CPU*

*[0] : Your GPU*

*[0] Which GPU indexes to choose? :*

**Pressing enter will use default value (0).** After that's done you will see a command line window with current settings as well as preview window which shows all the controls needed to operate the interactive converter/merger.

Here is a quick look at both the command line window and converter preview window:



Converter features many options that you can use to change the mask type, its size, feathering/blur, you can add additional color transfer and sharpen/enhance final trained face even further.

**Here is the list of all merger/converter features explained:**

### 1. Main overlay modes:

- original: displays original frame without swapped face
- overlay: simple overlays learned face over the frame
- hist-match: overlays the learned face and tries to match it based on histogram (has 2 modes: normal and masked that can be switched with Z)
- seamless: uses opencv poisson seamless clone function to blend new learned face over the head in the original frame- seamless hist match: combines both hist-match and seamless.
- raw-rgb: overlays raw learned face without any masking

**NOTE: Seamless modes can cause flickering.**

**2. Hist match threshold:** controls strength of the histogram matching in hist-match and seamless hist-match overlay mode.

*Q* - increases value

*A* - decreases value

**3. Erode mask:** controls the size of a mask.

*W - increases mask erosion (smaller mask)*

*S - decreases mask erosion (bigger mask)*

**4. Blur mask:** blurs/feathers the edge of the mask for smoother transition

*E - increases blur*

*D - decreases blur*

**5. Motion blur:** after entering initial parameters (converter mode, model, GPU/CPU) merger loads all frames and data\_dst aligned data, while it's doing it, it calculates motion vectors that are being used to create effect of motion blur which this setting controls, it let's you add it in places where face moves around but high values may blur the face even with small movement. The option only works if one set of faces is present in the "data\_dst/aligned" folder - if during cleanup you had some faces with \_1 prefixes (even if only faces of one person are present) the effect won't work, same goes if there is a mirror that reflects target persons face, in such case you cannot use motion blur and the only way to add it is to train each set of faces separately.

*R - increases motion blur*

*F - decreases motion blur*

**6. Super resolution:** uses similar algorithm as data\_src dataset/faceset enhancer, it can add some more definitions to areas such as teeth, eyes and enhance detail/texture of the learned face.

*T - increases the enhancement effect*

*G - decreases the enhancement effect*

**7. Blur/sharpen:** blurs or sharpens the learned face using box or gaussian method.

*Y - sharpens the face*

*H - blurs the face*

*N - box/gaussian mode switch*

**8. Face scale:** scales learned face to be larger or smaller.

*U - scales learned face down*

*J - scales learned face up*

**9. Mask modes:** there are 6 masking modes:

**dst:** uses masks derived from the shape of the landmarks generated during data\_dst faceset/dataset extraction.

**learned-prd:** uses masks learned during training. Keep shape of SRC faces.

**learned-dst:** uses masks learned during training. Keep shape of DST faces.

**learned-prd\*dst:** combines both masks, smaller size of both.

**learned-prd+dst:** combines both masks, bigger size of both.

**XSeg-prd:** uses XSeg model to mask using data from source faces.

**XSeg-dst:** uses XSeg model to mask using data from destination faces.

**XSeg-prd\*dst:** combines both masks, smaller size of both.

**learned-prd\*dst\*XSeg-dst\*prd:** combines all 4 mask modes, smaller size of all.

**10. Color transfer modes:** similar to color transfer during training, you can use this feature to better match skin color of the learned face to the original frame for more seamless and realistic face swap. There are 8 different modes:

**RCT**

**LCT**

**MKL**

**MKL-M**

**IDT**

**IDT-M**

**SOT - M**

**MIX-M**

**11. Image degrade modes:** there are 3 settings that you can use to affect the look of the original frame (without affecting the swapped face):

**Denoise** - denoises image making it slightly blurry (I - increases effect, K - decrease effect)

**Bicubic** - blurs the image using bicubic method (O - increases effect, L - decrease effect)

**Color** - decreases color bit depth (P - increases effect, ; - decrease effect)

**Additional controls:**

**TAB button** - switches between main preview window and help screen. For complete list of keys (and what they control, such as moving forward/backward, starting merging) check the help screen. Bear in mind these will only work in the main preview window, pressing any button while on the help screen won't do anything.

## 8. Conversion of frames back into video

After you merged/convert all the faces and you will have a folder named "merged" inside your "data\_dst" folder containing all frames as well as "merged\_masked" which contains mask frames.

Last step is to convert them back into video and combine with original audio track from data\_dst.mp4 file. To do so you will use one of 4 provided .bat files that will use FFmpeg to combine all the frames into a video in one of the following formats - avi, mp4, lossless mp4 or lossless mov:

- 8) merged to avi
- 8) merged to mov lossless
- 8) merged to mp4 lossless
- 8) merged to mp4

**And that's it!** After you've done all these steps you should have a file called result.xxx (avi/mp4/mov) which is your deepfake video as well as result\_mask.xxx which you can import into your video editing software to use as the source of masks to further refine your swapped face without affecting the rest of the video.