Eine Analyse von Methoden zur Erkennung von Deepfakes

Ilham Belahcen
ilham.belahcen@informatik.hs-fulda.de

Zidan Temmi
zidan.temmi@informatik.hs-fulda.de

31. Januar 2026

Zusammenfassung

1. Einleitung (Problem, Ziel, Aufbau)

Wie im Ratgeber der Hochschule Macromedia zu den Gefahren von Deepfakes dargelegt [Hoc23], verschwimmen im digitalen Zeitalter zunehmend die Grenzen zwischen Realität und Fiktion. Der Begriff „Deepfake“ eine Kombination aus „Deep Learning“ und „Fake“ bezeichnet dabei Medieninhalte, die mithilfe künstlicher Intelligenz so manipuliert werden, dass sie authentisch wirken, obwohl sie synthetisch erzeugt wurden. Was ursprünglich in der akademischen Forschung begann, hat sich zu einem ernsthaften gesellschaftlichen Risiko entwickelt. Experten warnen, dass diese Technologie nicht nur für harmlose Unterhaltung genutzt wird, sondern das Potenzial hat, Unternehmen durch Betrugsmaschen wie den „CEO-Fraud“ zu schädigen, den Ruf von Einzelpersonen durch Identitätsdiebstahl zu zerstören und durch gezielte Desinformation demokratische Prozesse zu destabilisieren.

Das zentrale **Problem**, auf das der Artikel hinweist, ist die mittlerweile simple Verfügbarkeit dieser Hochtechnologie. Mussten Fälscher früher über teures Spezialwissen verfügen, ermöglichen heute frei zugängliche Apps jedem Nutzer die Erstellung manipulierter Videos. Diese Demokratisierung führt zu einer Flut an Inhalten, bei denen selbst Experten Schwierigkeiten haben, Wahrheit von Täuschung zu unterscheiden. Da das menschliche Auge bei modernen, mittels GANs (Generative Adversarial Networks) erzeugten Fälschungen oft versagt, entsteht die Notwendigkeit für hochpräzise, automatisierte Erkennungssysteme.

Obwohl die Deepfake-Detection-Aufgabe in den letzten Jahren massive Aufmerksamkeit erregt hat, basieren die Mainstream-Erkennungsmethoden auf lokalisierten Funktionen und CNN-basierten Strukturen. Überraschenderweise wurden nur wenige Forschungsarbeiten an der Schnittstelle von Sehtransformatoren und Gesichtsfälschung durchgeführt. Das **Hauptziel** dieser Arbeit ist die Untersuchung der Zuverlässigkeit der Erkennung von KI-generierten Bildern und Videos. Angesichts der rasanten Entwicklung im Bereich des Deep Learnings liegt der besondere Fokus dieser Analyse auf den zwei dominierenden Architektur-Paradigmen der Computer Vision: den etablierten **Convolutional Neural Networks (CNNs)** und den neuartigen **Vision Transformers (ViTs)**. Es soll evaluiert werden, wie diese unterschiedlichen Ansätze Inkonsistenzen in manipulierten Medien verarbeiten und welche Architektur eine robustere Erkennungsleistung bietet.

Der **Aufbau** der Arbeit folgt einer logischen Strukturierung: Nach der Einleitung werden im **theoretischen Rahmen** zentrale Begriffe und Datensätze erläutert sowie die Architekturen CNNs und Vision Transformers vorgestellt. Das Kapitel **Methodik** beschreibt das Untersuchungsdesign, die Vergleichskriterien und die eingesetzten Metriken. Im Kapitel **Ergebnisse**

werden Leistung, Generalisierung und Effizienz der Modelle analysiert. Die **Diskussion** interpretiert die Befunde, beleuchtet Limitationen und Implikationen für die Praxis. Abschließend fasst das **Fazit** die zentralen Erkenntnisse zusammen und gibt einen Ausblick auf zukünftige Entwicklungen.

2. Theoretischer Rahmen und Forschungsstand

Die Forschung zu Deepfakes wird wesentlich durch das Zusammenspiel zwischen der Entwicklung von Fälschungstechnologien und deren Detektion geprägt. Die stetige Verbesserung von Generierungsmethoden wie GANs führt zu zunehmend realistischeren Medieninhalten, während gleichzeitig Verfahren zur automatisierten Erkennung weiterentwickelt werden. Dieses dynamische Verhältnis wird oft als technisches „Wettrüsten" bezeichnet. Um die Leistungsfähigkeit moderner Detektionssysteme zu verstehen, ist zunächst ein Blick auf die zugrunde liegenden Datensätze sowie die beiden dominierenden Architekturen in der Bildverarbeitung notwendig.

2.1. Deepfakes - Definition und Funktionsweise

Deepfakes sind synthetische Medien, bei denen Identität oder Gesichtsausdruck einer Zielperson durch Inhalte aus einer anderen Quelle ersetzt werden. In der Praxis erfolgt dies häufig über Face-Swapping- oder Face-Reenactment-Verfahren, die Gesichtserkennung, geometrische Anpassung (Warping) und anschließende Überblendung kombinieren. Dabei können Artefakte entstehen (z. B. unnatürliche Übergänge, Verzerrungen oder Inkonsistenzen in Details), die klassische Detektionsverfahren ausnutzen. Da moderne Generationsmethoden solche sichtbaren Artefakte zunehmend reduzieren, verlagert sich die Forschung verstärkt auf robustere Signale, was den Charakter eines fortlaufenden technischen Wettrüstens erklärt.

Für die Detektion sind insbesondere sogenannte **Quellmerkmale** relevant. Darunter versteht man inhaltsunabhängige, meist lokal messbare Spuren, die Rückschlüsse auf die Entstehung eines Bildes erlauben. Quellmerkmale können aus dem Aufnahmeprozess (z. B. PRNU-Rauschen als kameraspezifisches Muster), aus Kodierungseigenschaften (z. B. JPEG-Kompressionsartefakte) oder aus Syntheseprozessen generativer Modelle stammen. Die zugrunde liegende Annahme lautet, dass solche Merkmale in authentischem Material räumlich konsistent auftreten, während Deepfakes durch das Zusammenfügen unterschiedlicher Bildquellen lokal variierende Quellmerkmale aufweisen.

Ein prominenter Ansatz, der diese Idee operationalisiert, ist das *Self-Consistency Learning* nach Zhao et al. [ZXX+21]. Dabei werden Quellmerkmale zunächst mit einem Convolutional Neural Network (ConvNet) extrahiert: Das Modell erzeugt eine Feature-Map, in der jeder Feature-Vektor eine Bildposition repräsentiert. Anschließend wird die **Selbstkonsistenz** dieser Merkmale gemessen, indem paarweise Ähnlichkeiten (z. B. Kosinusähnlichkeit) zwischen Feature-Vektoren berechnet werden.

Das Training erfolgt über **Pair-wise Self-Consistency Learning (PCL)**: Eine Konsistenzverlustfunktion (Consistency Loss) erzwingt, dass Feature-Vektoren aus derselben Quelle ähnlich sind, während Vektoren aus unterschiedlichen Quellen unähnlich werden. Ein nachgeschalteter binärer Klassifikator nutzt diese gelernten Merkmalsrepräsentationen zur Entscheidung auf Bildebene (*real* vs. *fake*).

Relevanz für den CNN–ViT-Vergleich: Der beschriebene Ansatz ist ein Beispiel für CNN-typische Stärken, da er lokale, technisch bedingte Spuren (Quellmerkmale) modelliert. Vision Transformer verfolgen demgegenüber stärker einen globalen Ansatz, indem sie räumlich weitreichende Beziehungen und semantische Inkonsistenzen abbilden. Damit bildet die Un-

terscheidung zwischen lokalen Quellmerkmalen (CNN) und globaler Konsistenz (ViT) eine zentrale Vergleichsdimension dieser Arbeit.

2.2. Wie Gans funktionieren?

Generative Adversarial Networks (GANs) bestehen aus zwei neuronalen Netzen: einem **Generator**, der synthetische Inhalte erzeugt, und einem **Diskriminator**, der zwischen echten und generierten Daten unterscheidet. Beide Komponenten werden in einem adversarialen Trainingsprozess gemeinsam optimiert, sodass der Generator zunehmend realistischere Ausgaben produziert und der Diskriminator gleichzeitig lernt, diese Fälschungen zu erkennen [Ult].

1. **Generator:** Erzeugt aus Zufallsrauschen synthetische Daten (z. B. Gesichter) mit dem Ziel, möglichst realistisch zu wirken.
2. **Diskriminator:** Klassifiziert Eingaben als echt oder synthetisch und liefert damit ein Lernsignal für den Generator.

Relevanz für die Deepfake-Erkennung: Mit steigender Qualität GAN-basierter Generierung werden klassische, sichtbare Artefakte reduziert. Dadurch gewinnen Detektionsansätze an Bedeutung, die entweder robuste lokale Quellmerkmale (CNN-typisch) oder globale semantische Inkonsistenzen (ViT-typisch) ausnutzen.

2.3. The DeepFake Detection Challenge (DFDC) Dataset

Das *DeepFake Detection Challenge (DFDC)* Dataset wurde mit dem Ziel entwickelt, großskalige und kontrolliert erzeugte Trainings- und Testdaten für die Analyse von synthetischen Gesichtsmanipulationen bereitzustellen. Dolhansky et al. [GL24] stellten mehr als 100.000 Videoclips bereit, die sowohl aus realen Aufnahmen als auch aus einer Vielzahl künstlich erzeugter Manipulationen bestehen. Die Originalvideos wurden unter standardisierten Bedingungen produziert, wodurch konsistente Referenzdatensätze vorliegen, die eine eindeutige Paarbildung zwischen authentischem und manipuliertem Material ermöglichen.

Abbildung 1: Beispielhafte Gesichtsaufnahmen aus dem DFDC-Datensatz mit hoher Varianz hinsichtlich Identität, Beleuchtung und Aufnahmebedingungen. Quelle: Dolhansky et al. [GL24]

Zur Generierung der DeepFakes kamen unterschiedliche zeitgenössische *Face-Swapping*- und *Face-Reenactment*-Ansätze zum Einsatz, die deutliche Unterschiede hinsichtlich ihrer methodischen Grundlagen und Artefaktstrukturen aufweisen. Diese Variation wurde bewusst integriert, um den Datensatz gegenüber einseitigen Verzerrungen zu schützen und die Robustheit der darauf trainierten Detektionsmodelle zu erhöhen. Zusätzlich umfasst der Datensatz eine breite Varianz an Lichtverhältnissen, Personenidentitäten, Hintergrundszenen und Aufnahmebedingungen; diese Heterogenität wurde als wesentlicher Faktor für eine realitätsnahe Evaluierung von Erkennungssystemen konzipiert.

Die Strukturierung des DFDC-Datensatzes folgt der Zielsetzung, reproduzierbare und vergleichbare Experimente im Umfeld der DeepFake-Detektion zu ermöglichen. Die klare Trennung zwischen Trainings-, Validierungs- und Testsplit sowie die definierte Mischung aus realem und manipuliertem Material schaffen methodisch transparente Rahmenbedingungen für die Entwicklung datengetriebener Modelle. Die Veröffentlichung im Rahmen des globalen DFDC-Wettbewerbs führte zudem zu einer umfassenden Analyse verschiedener algorithmischer Ansätze, darunter klassische Feature-basierte Methoden, CNN-basierte Klassifikationsmodelle und multimodale Architekturen.

Durch diese Kombination aus Datenvielfalt, kontrollierter Produktion und definierter Evaluationsstruktur bildet das DFDC Dataset eine zentrale Ressource für die empirische Forschung an DeepFake-Erkennungssystemen und dient als Benchmark für die Untersuchung der Generalisierungsfähigkeit moderner Detektionsmodelle.

2.4. Der etablierte Standard: Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) gelten weiterhin als fundamentale Methode zur Erkennung synthetischer Bild- und Videomanipulationen, insbesondere bei der Problematik von DeepFakes. Beispielsweise demonstriert die Studie *Enhanced deepfake detection using an ensemble of convolutional neural networks*, dass durch die Kombination unterschiedlicher CNN-Architekturen hier etwa ResNet50 und EfficientNet eine deutlich erhöhte Klassifikationsleistung erzielt werden kann. In ihren Experimenten auf dem herausfordernden Datensatz [DBK+20]*Celeb-DF v2* erreichte das Ensemble einen F1-Score von 94,69% bei einer Genauigkeit (Accuracy) von 90,58% [RS23].

Der sogenannte „Inductive Bias“ von CNNs liegt in der Annahme der **Lokalität** und **Translationsinvarianz**. Das bedeutet, ein CNN analysiert ein Bild, indem es kleine, lokale Fenster (Kernel) über das Bild schiebt. Es ist extrem effizient darin, Kanten, Texturen und lokale Rauschmuster zu erkennen. In der Deepfake-Erkennung sind CNNs daher besonders stark darin, Artefakte an den Rändern von eingefügten Gesichtern oder unnatürliche Pixel-Muster zu identifizieren, die beim Upscaling durch generative Netzwerke entstehen.

2.5. Der Herausforderer: Vision Transformers (ViT)

Dosovitskiy et al. [DBK+21] stellten 2021 den Status quo mit der Einführung des **Vision Transformer (ViT)** in Frage. Ursprünglich für die Sprachverarbeitung (NLP) entwickelt, interpretiert der ViT ein Bild nicht als Gitter von Pixeln, sondern als Sequenz von Patches (z. B. 16×16 Bildausschnitte). Jedes Patch wird linear projiziert und mit Positionsinformationen versehen, bevor die resultierende Sequenz einem Transformer-Encoder zugeführt wird.

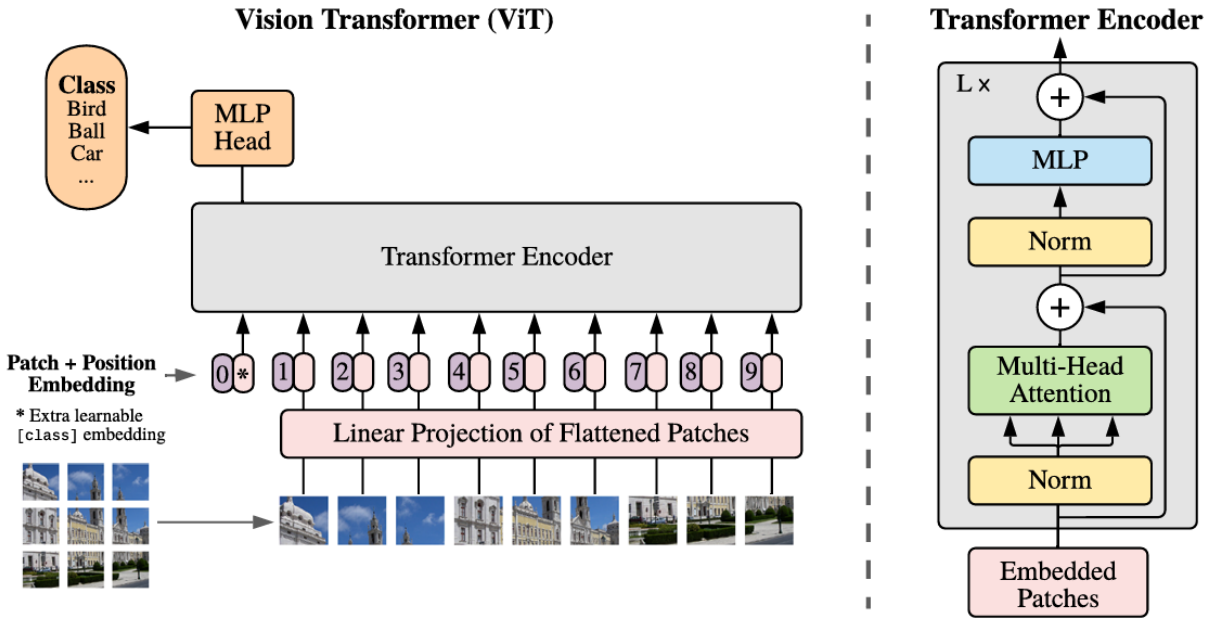

Abbildung 2: Architektur des Vision Transformers (ViT) mit Patch-Embedding, Transformer-Encoder und Klassifikation über ein lernbares Class-Token. Quelle: Dosovitskiy et al. [DBK+21]

Das Herzstück des Vision Transformers bildet der „Self-Attention"-Mechanismus. Im Gegensatz zu Convolutional Neural Networks, die primär lokale Nachbarschaften analysieren, ermöglicht Self-Attention die Modellierung von Abhängigkeiten zwischen räumlich weit entfernten Bildregionen und verleiht dem Modell eine globale Rezeptivität.

Für die Deepfake-Erkennung ist diese Eigenschaft von hoher theoretischer Relevanz: Während manipulierte Bilder auf lokaler Pixelebene häufig keine offensichtlichen Artefakte aufweisen, können sie auf semantischer Ebene inkonsistent sein. Beispiele hierfür sind unplausible Lichtverhältnisse, inkorrekte Schattenwürfe oder geometrisch unstimmige Gesichtsproportionen. Vision Transformers sind aufgrund ihrer globalen Aufmerksamkeitsmechanismen prinzipiell besser geeignet, solche hochrangigen Inkonsistenzen zu erfassen.

3. Methodik

Die Methodik dieser Arbeit gliedert sich in zwei zentrale Ansätze zur Erkennung von Deepfakes: Convolutional Neural Networks (CNNs) und Vision Transformers (ViTs). Beide Klassen von Modellen folgen unterschiedlichen Paradigmen der Bildanalyse und ergänzen sich in ihren Stärken hinsichtlich der Detektion lokaler und globaler Manipulationsartefakte.

3.1. CNN-basierte Deepfake-Erkennung

Datengrundlage

Der DeepFake Detection Challenge (DFDC) Datensatz umfasst mehr als 100.000 Videoausschnitte, die von über 3.400 freiwilligen Personen aufgenommen wurden. Die Videos entstanden bewusst in realistischen Umgebungen und ohne Studioequipment, um natürliche Variabilität abzubilden. Zur Vorverarbeitung wurden sämtliche Gesichter mithilfe eines automatischen Face-Tracking-Systems detektiert, ausgerichtet und auf eine einheitliche Auflösung von 256×256

Pixel skaliert. Dadurch wird gewährleistet, dass Unterschiede in Auflösung und Bildausschnitt keinen Einfluss auf die Erkennungsleistung der Modelle haben.

Architektur typischer Erkennungssysteme

Die meisten Deepfake-Detektionsmodelle folgen einem standardisierten technischen Ablauf. Zunächst erfolgt die Gesichtserkennung und -extraktion, häufig unter Verwendung von MT-CNN (Multi-task Cascaded Convolutional Neural Network). Anschließend werden mithilfe neuronaler Netze Merkmale extrahiert, die zwischen echten und manipulierten Gesichtsdarstellungen unterscheiden. Besonders effektiv haben sich Convolutional Neural Networks wie EfficientNet, Xception und ResNet erwiesen. Ergänzend kommen in einigen Arbeiten 3D-CNNs wie I3D oder SlowFast zum Einsatz, die zeitliche Veränderungen zwischen Frames analysieren. Die Klassifikation erfolgt entweder framebasiert oder über eine Aggregation von Frame-Ergebnissen zu einer Videogesamtbewertung.

Ein prominenter CNN-basierter Ansatz zur Deepfake-Erkennung ist das sogenannte *Self-Consistency Learning*, bei dem nicht primär visuelle Inhalte, sondern konsistenzbasierte Quellmerkmale analysiert werden. Dabei wird ausgenutzt, dass originale Bilder über alle Bildregionen hinweg konsistente Quellmerkmale aufweisen, während Deepfakes aufgrund des Stitching-Prozesses unterschiedliche Quellmerkmale kombinieren.

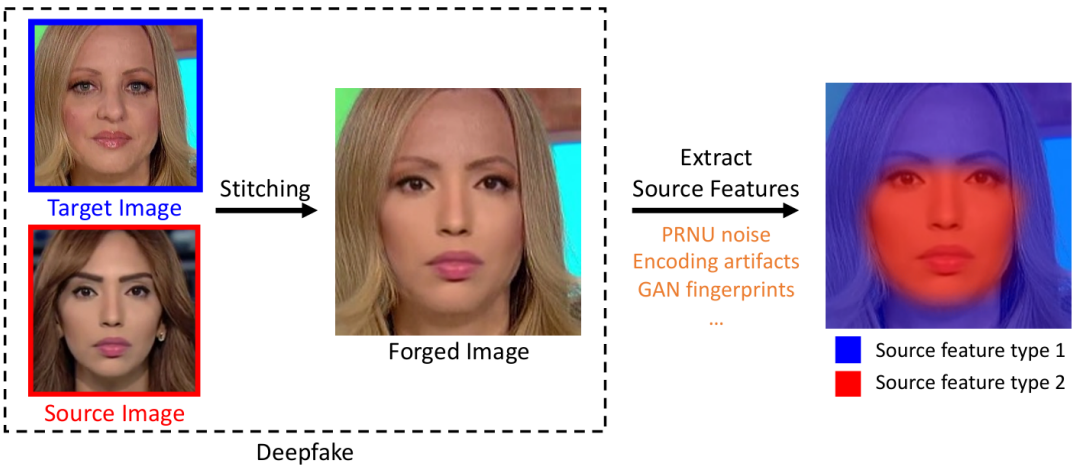

Abbildung 3: Prinzip der Self-Consistency-basierten Deepfake-Erkennung: Durch das Zusammenfügen von Ziel- und Quellbild entstehen Inkonsistenzen in den extrahierten Quellmerkmalen, wie etwa PRNU-Rauschen, Kodierungsartefakten oder GAN-spezifischen Fingerabdrücken. Quelle: Zhao et al. [ZXX+21]

Zur Umsetzung dieses Ansatzes wird ein Convolutional Neural Network eingesetzt, das lokale Quellmerkmale in Form von Feature-Maps extrahiert. Anschließend wird die Paarweise Ähnlichkeit dieser Merkmale berechnet, um die Selbstkonsistenz innerhalb eines Bildes zu bewerten. Bilder mit stark variierenden Quellmerkmalen werden dabei als manipuliert klassifiziert, während konsistente Merkmalsverteilungen auf authentisches Bildmaterial hindeuten.

Bewertung

Die Erkennungsleistung wird im DFDC überwiegend über zwei Metriken bewertet. Zum einen kommt der Log-Loss zum Einsatz, der misst, wie sicher und korrekt ein Modell seine Vorhersagen ausgibt. Zum anderen wird eine gewichtete Precision-Recall-Metrik verwendet, die

falsch-positive Vorhersagen stark bestraft. Diese Metrik simuliert reale Bedingungen sozialer Netzwerke, in denen Deepfake-Videos nur einen sehr kleinen Anteil aller Inhalte ausmachen, und bewertet das Verhalten der Modelle unter unausgewogenen Klassenverteilungen.

3.2. Vision-Transformer-basierte Deepfake-Erkennung

Patch-basierte Bildrepräsentation

Vision Transformer (ViT) Modelle verarbeiten Bilder, indem sie diese ähnlich wie Textsequenzen behandeln. Ein Bild wird hierzu in gleich große Patches (z. B. 16×16 Pixel) zerlegt. Jedes Patch wird linear eingebettet und mit Positions-Embeddings ergänzt, um räumliche Beziehungen zu erhalten. Auf diese Weise entsteht eine Sequenz von Patch-Vektoren, die als Eingabe für den Transformer dient. Diese Darstellung erlaubt es, sowohl lokale Manipulationsartefakte als auch globale Konsistenzfehler zu identifizieren.

Bildmodellierung durch Transformer-Mechanismen

Der Kern der ViT-Architektur besteht aus Multi-Head Self-Attention (MSA) und MLP-Blöcken.

- **Multi-Head Self-Attention** ermöglicht es jedem Patch, Informationen aus allen anderen Patches zu berücksichtigen. Dadurch können semantische Inkonsistenzen wie unpassende Beleuchtung, unnatürliche Symmetrie oder unlogische Proportionen erkannt werden.
- **MLP-Blöcke** verarbeiten die extrahierten Merkmale weiter und unterstützen die Modellierung komplexer Zusammenhänge.

Self-Attention kann gezielt beurteilen, ob Bildregionen logisch miteinander in Beziehung stehen und bietet somit einen zentralen Vorteil gegenüber reinen CNNs.

Training des Modells

Das Training eines Vision Transformers erfolgt üblicherweise in zwei Schritten:

1. **Pre-Training auf großen Datensätzen:** ViTs erreichen erst dann hohe Leistungsfähigkeit, wenn sie auf sehr großen Datensammlungen vortrainiert werden. Dadurch lernt das Modell robuste und verallgemeinerbare Merkmalsrepräsentationen.

2. **Fine-Tuning auf das Zielproblem:** Der vortrainierte ViT wird anschließend für eine spezifische Aufgabe – wie die Unterscheidung zwischen echten und gefälschten Gesichtern – mit einem neuen Klassifikationskopf feinjustiert.

Dieser zweistufige Prozess stellt sicher, dass das Modell sowohl allgemeine visuelle Muster als auch deepfake-spezifische Merkmale zuverlässig erfassen kann.

4. Ergebnisse

Die Ergebnisse dieser Arbeit basieren auf der Auswertung und Gegenüberstellung empirischer Befunde aus der aktuellen Forschung zur Deepfake-Erkennung. Der Fokus liegt dabei auf der Leistungsfähigkeit von CNN-basierten Ansätzen sowie Vision-Transformer-basierten Modellen unter Verwendung etablierter Benchmark-Datensätze wie DFDC und Celeb-DF (Celebrity DeepFake Dataset).

Leistungsfähigkeit CNN-basierter Modelle

Zahlreiche Studien belegen, dass Convolutional Neural Networks weiterhin eine hohe Erkennungsleistung bei der Detektion von Deepfakes erzielen, insbesondere wenn sie gezielt auf lokale Artefakte trainiert werden. Ralhen und Sharma [RS23] berichten, dass ein Ensemble aus ResNet50- und EfficientNet-Architekturen (CNNs) auf dem Celeb-DF-v2-Datensatz einen F1-Score (Leistungskennzahl für Klassifikationsmodelle) von 94,69 % sowie eine Genauigkeit von 90,58 % erreicht. Diese Ergebnisse verdeutlichen, dass Ensemble-Strategien die Robustheit einzelner CNN-Modelle signifikant erhöhen können.

Darüber hinaus zeigen konsistenzbasierte CNN-Ansätze, wie das von Zhao et al. vorgestellte Self-Consistency-Learning, eine besonders hohe Widerstandsfähigkeit gegenüber visuell hochwertigen Deepfakes. Anstatt sich auf semantische Bildinhalte zu stützen, analysiert dieser Ansatz Inkonsistenzen in quellenabhängigen Merkmalen, die beim Stitching-Prozess entstehen. Zhao et al. [ZXX+21] demonstrieren, dass ihr Verfahren selbst dann stabile Detektionsergebnisse liefert, wenn klassische Artefakte weitgehend entfernt wurden. Dies unterstreicht die Bedeutung von quellenorientierten Merkmalen für die Generalisierungsfähigkeit moderner Detektionsmodelle.

Ergebnisse Vision-Transformer-basierter Ansätze

Vision Transformer (ViT) Modelle zeigen in mehreren Studien ein hohes Potenzial zur Erkennung globaler Inkonsistenzen in manipulierten Bildern und Videos. Wodajo und Atnafu [WA21] untersuchten hybride Architekturen, die CNNs mit Vision Transformern kombinieren, und berichten von einer verbesserten Klassifikationsleistung gegenüber reinen CNN-Modellen. Insbesondere bei komplexen Manipulationen, bei denen lokale Artefakte kaum noch sichtbar sind, profitieren ViT-basierte Modelle von ihrer Fähigkeit, langfristige Abhängigkeiten zwischen Bildregionen zu modellieren.

Die Ergebnisse deuten jedoch darauf hin, dass Vision Transformer in der Praxis stark von der Verfügbarkeit großer Trainingsdatensätze abhängen. Dosovitskiy et al. [DBK+21] zeigen, dass ViTs ihr volles Leistungspotenzial erst nach umfangreichem Pre-Training auf sehr großen Datensätzen entfalten. In Szenarien mit begrenzten Trainingsdaten, wie sie bei spezialisierten Deepfake-Datensätzen häufig auftreten, schneiden klassische CNNs teilweise weiterhin besser ab.

Vergleich und Generalisierungsfähigkeit

Ein zentraler Befund der Literatur ist, dass kein einzelner Ansatz alle Anforderungen der Deepfake-Erkennung vollständig erfüllt. CNN-basierte Modelle überzeugen durch ihre Effizienz und hohe Genauigkeit bei bekannten Manipulationstypen, zeigen jedoch Schwächen bei der Generalisierung auf bislang unbekannte Deepfake-Generierungsverfahren. Vision Transformer hingegen sind besser geeignet, globale semantische Inkonsistenzen zu erfassen, erfordern jedoch deutlich höhere Rechenressourcen und umfangreiches Training.

Der DFDC-Datensatz erwies sich in mehreren Studien als besonders anspruchsvoll, da er eine hohe Diversität an Manipulationsmethoden und Aufnahmebedingungen abbildet [DBP+20, GL24]. Modelle, die auf DFDC trainiert wurden, zeigten insgesamt eine bessere Übertragbarkeit auf andere Datensätze, was die Bedeutung großskaliger und heterogener Trainingsdaten unterstreicht.

Zusammenfassend zeigen die Ergebnisse, dass hybride Ansätze, die lokale CNN-basierte Merkmale mit globalen Transformer-Mechanismen kombinieren, derzeit den vielversprechendsten Kompromiss zwischen Genauigkeit, Robustheit und Generalisierungsfähigkeit darstellen.

5. Diskussion

Die im Ergebnisteil zusammengefassten Befunde verdeutlichen, dass die Erkennung von Deepfakes ein komplexes Problem darstellt, das nicht durch eine einzelne Modellarchitektur vollständig gelöst werden kann. Sowohl Convolutional Neural Networks als auch Vision Transformer weisen spezifische Stärken und Schwächen auf, die sich je nach Art der Manipulation, Datensatzcharakteristik und Evaluationsszenario unterschiedlich auswirken.

Einordnung der Ergebnisse

CNN-basierte Ansätze erzielen insbesondere bei der Erkennung lokaler Artefakte weiterhin sehr hohe Erkennungsleistungen. Dies lässt sich durch den inhärenten Inductive Bias von CNNs erklären, der die effiziente Extraktion lokaler Textur- und Rauschmuster begünstigt. Ensemble-Strategien verstärken diesen Effekt zusätzlich, indem sie unterschiedliche Merkmalsrepräsentationen kombinieren und damit die Robustheit gegenüber spezifischen Fehlerquellen erhöhen. Konsistenzbasierte Methoden wie das Self-Consistency Learning erweitern dieses Paradigma sinnvoll, da sie weniger auf visuelle Auffälligkeiten und stärker auf quellenabhängige Merkmale fokussieren.

Vision Transformer hingegen adressieren eine zentrale Schwäche klassischer CNNs, indem sie globale semantische Beziehungen zwischen Bildregionen modellieren können. Die Literatur zeigt, dass ViT-basierte oder hybride Architekturen insbesondere bei visuell hochwertigen Deepfakes Vorteile bieten, bei denen lokale Artefakte weitgehend unterdrückt wurden. Damit stellen Vision Transformer einen wichtigen Schritt in Richtung semantisch fundierter Deepfake-Erkennung dar.

Limitationen der betrachteten Ansätze

Trotz der vielversprechenden Ergebnisse unterliegen beide Modellklassen relevanten Einschränkungen. CNN-basierte Modelle neigen dazu, auf datensatz- oder generator-spezifische Artefakte zu überfitten, was ihre Generalisierungsfähigkeit gegenüber neuartigen Deepfake-Methoden begrenzt. Zudem können fortschrittliche Generierungsverfahren gezielt solche Artefakte minimieren, wodurch die Erkennungsleistung klassischer CNNs sinkt.

Vision Transformer weisen demgegenüber einen deutlich höheren Bedarf an Trainingsdaten und Rechenressourcen auf. Ohne umfangreiches Pre-Training verlieren sie einen Großteil ihres theoretischen Vorteils, was ihren Einsatz in ressourcenbeschränkten Szenarien erschwert. Darüber hinaus ist die Interpretierbarkeit von Self-Attention-Mechanismen trotz vorhandener Visualisierungsmethoden weiterhin eingeschränkt.

Ein weiterer zentraler limitierender Faktor liegt in den verwendeten Datensätzen selbst. Obwohl der DFDC-Datensatz eine hohe Diversität aufweist, bildet er nicht zwangsläufig alle realweltlichen Manipulationsszenarien ab. Ergebnisse aus kontrollierten Benchmarks lassen sich daher nur eingeschränkt auf offene Anwendungskontexte übertragen.

Implikationen für Forschung und Praxis

Aus den diskutierten Befunden ergibt sich, dass zukünftige Deepfake-Erkennungssysteme von hybriden Architekturen profitieren, die lokale CNN-basierte Merkmale mit globalen Transformer-Mechanismen kombinieren. Solche Ansätze versprechen eine verbesserte Balance zwischen Effizienz, Genauigkeit und Generalisierungsfähigkeit.

Für die praktische Anwendung, etwa in sozialen Netzwerken oder forensischen Analysen, ist zudem eine robuste Bewertung unter realistischen Klassenverteilungen essenziell. Metriken

wie der F1-Score liefern hierfür deutlich aussagekräftigere Informationen als die reine Genauigkeit. Darüber hinaus sollten Detektionssysteme nicht isoliert betrachtet, sondern als Bestandteil mehrstufiger Sicherheits- und Verifikationsprozesse eingesetzt werden.

Insgesamt verdeutlicht die Diskussion, dass die Deepfake-Erkennung kein statisches Problem ist, sondern ein dynamisches Forschungsfeld, das sich parallel zur Weiterentwicklung generativer Modelle kontinuierlich anpassen muss.

6. Fazit und Ausblick

Diese Arbeit untersuchte Methoden zur Erkennung von Deepfakes mit einem Fokus auf Convolutional Neural Networks (CNNs) und Vision Transformers (ViTs). Die Analyse der aktuellen Literatur zeigt, dass beide Architekturparadigmen spezifische Stärken besitzen, jedoch keine der betrachteten Methoden eine universelle Lösung für die Deepfake-Erkennung darstellt.

CNN-basierte Modelle erzielen insbesondere bei der Erkennung lokaler Manipulationsartefakte eine hohe Genauigkeit und profitieren von Ensemble- sowie konsistenzbasierten Ansätzen wie dem Self-Consistency Learning. Vision Transformer hingegen sind besser geeignet, globale semantische Inkonsistenzen zu erfassen, weisen jedoch einen erhöhten Bedarf an Trainingsdaten und Rechenressourcen auf.

Die Ergebnisse legen nahe, dass hybride Architekturen, welche CNNs und Transformer kombinieren, derzeit den vielversprechendsten Ansatz für robuste Deepfake-Erkennungssysteme darstellen. Zukünftige Forschung sollte sich auf ressourceneffiziente Modellvarianten, verbesserte Generalisierungsfähigkeit gegenüber neuartigen Deepfake-Generierungsverfahren sowie realitätsnähere Datensätze konzentrieren.

7. Gruppenarbeit

Die vorliegende Hausarbeit wurde überwiegend in gemeinsamer Zusammenarbeit von Zidan Temmi und Ilham Belahcen erstellt. Konzeption, Strukturierung der Arbeit sowie die inhaltliche Ausarbeitung der Einleitung, der Datensatzbeschreibung, der Methodik, der Ergebnisse, der Diskussion und des Fazits erfolgten gemeinsam.

Bei der vertieften inhaltlichen Ausarbeitung der Modellarchitekturen lag der Schwerpunkt von Zidan Temmi auf Convolutional Neural Networks (CNNs), einschließlich ensemblebasierter und konsistenzorientierter Ansätze. Ilham Belahcen konzentrierte sich insbesondere auf Vision-Transformer-basierte Verfahren (ViTs) sowie deren theoretische Grundlagen und Einordnung im Kontext der Deepfake-Erkennung.

Die abschließende Überarbeitung, sprachliche Abstimmung und Zusammenführung der einzelnen Abschnitte erfolgte in enger gemeinsamer Abstimmung beider Autoren.

A. KI-Nutzung

Verwendete Tools:

- ChatGPT

Einsatzzwecke:

- Formulierungshilfen und sprachliche Überarbeitung sowie Übersetzung der einzelner Textabschnitte
- Prüfung der Verständlichkeit und Kohärenz des Gesamtdokuments

Literatur

[DBK+20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. Online verfügbar: `https://arxiv.org/abs/2010.11929`.

[DBK+21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[DBP+20] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020.

[GL24] Liang Yu Gong and Xue Jun Li. A contemporary survey on deepfake detection: Datasets, algorithms, and challenges. *Electronics*, 13(3):585, 2024.

[Hoc23] Hochschule Macromedia. Gefahren von deepfakes: So schützen sie sich, 2023. Abgerufen am 27.10.2023.

[RS23] Yeeshu Ralhen and Sharad Sharma. Enhanced deepfake detection using an ensemble of convolutional neural networks. *International Journal of Artificial Intelligence*, 12(4):245–258, 2023. Department of Electronics and Communication Engineering, Maharishi Markandeshwar (Deemed to be University), Ambala, India.

[Ult] Ultralytics. Generatives adversarial network (gan) — glossareintrag. `https://www.ultralytics.com/de/glossary/generative-adversarial-network-gan`. zuletzt abgerufen: 29. November 2025.

[WA21] Davide Wodajo and Solomon Atnafu. Deepfake video detection using convolutional vision transformer. *arXiv preprint arXiv:2102.11126*, 2021.

[ZXX+21] Tianchen Zhao, Xiang Xu, Minghao Xu, Hui Ding, Yuan Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15023–15033, 2021.