Dr. Sascha Zmudzinski

Verteidigungsmaßnahmen zu Video-Deepfakes

Fraunhofer SIT

Schädliche Verwendung von Video-Deepfakes

Ansätze zum Erkennen von Video-Deepfakes

Manuelle, visuelle Betrachtung

Traditionelle Bildforensik

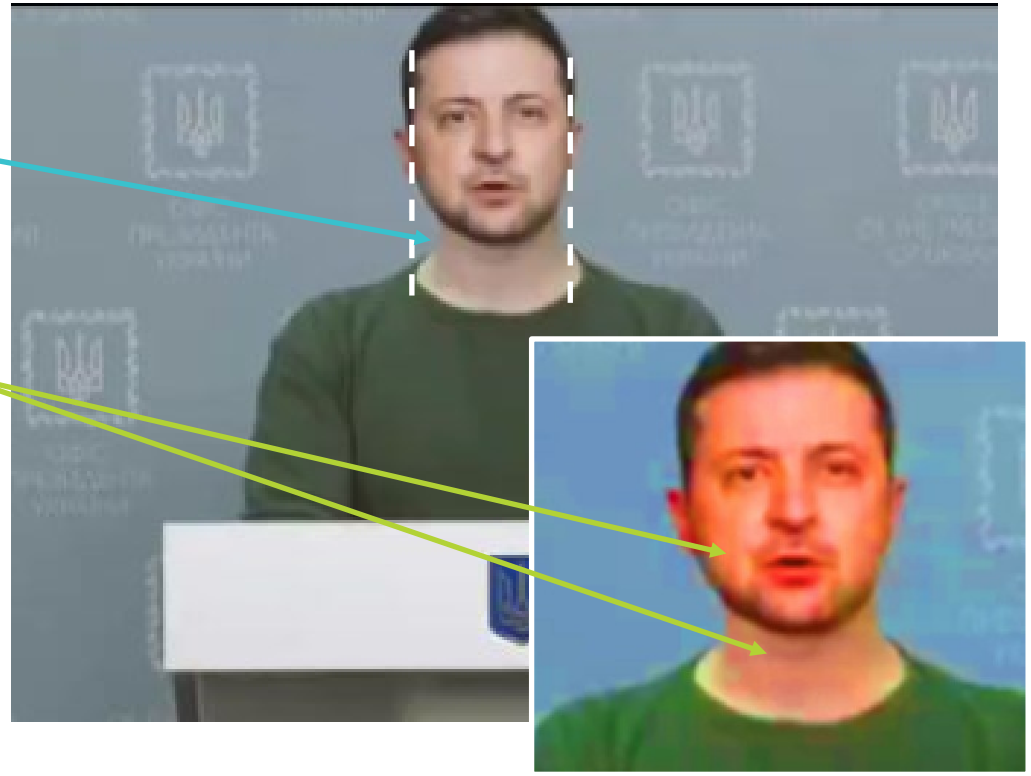
Wiedererkennen des Bildhintergrundes

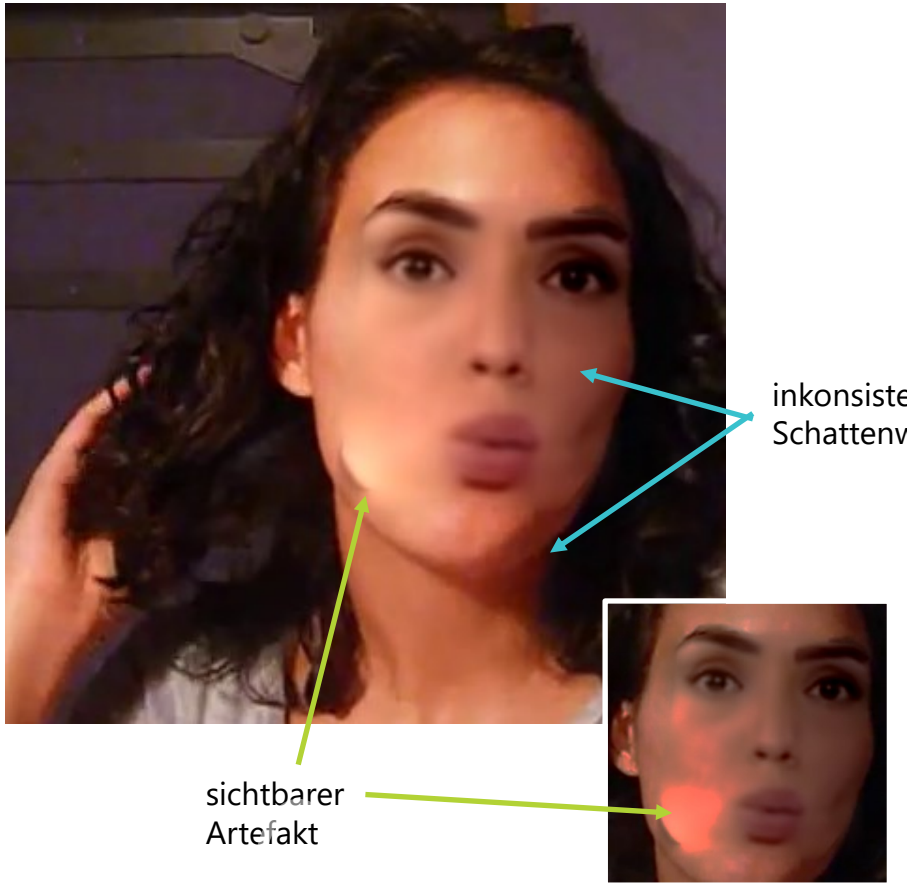
KI-basierte Erkennungsmethoden

© Fraunhofer

Deepfake Erkennung in Gesichtsregion

Manuelle, visuelle Betrachtung





Ansatz: sichtbare Auffälligkeiten manuell-visuell erkennen

- ggf. manuelle Analyse in Bildverarbeitungssoftware

© Fraunhofer

Quellen:
Youtube https://www.youtu.be/X17yrEV5sl4 (17.03.2022)
FaceForensics++ Dataset

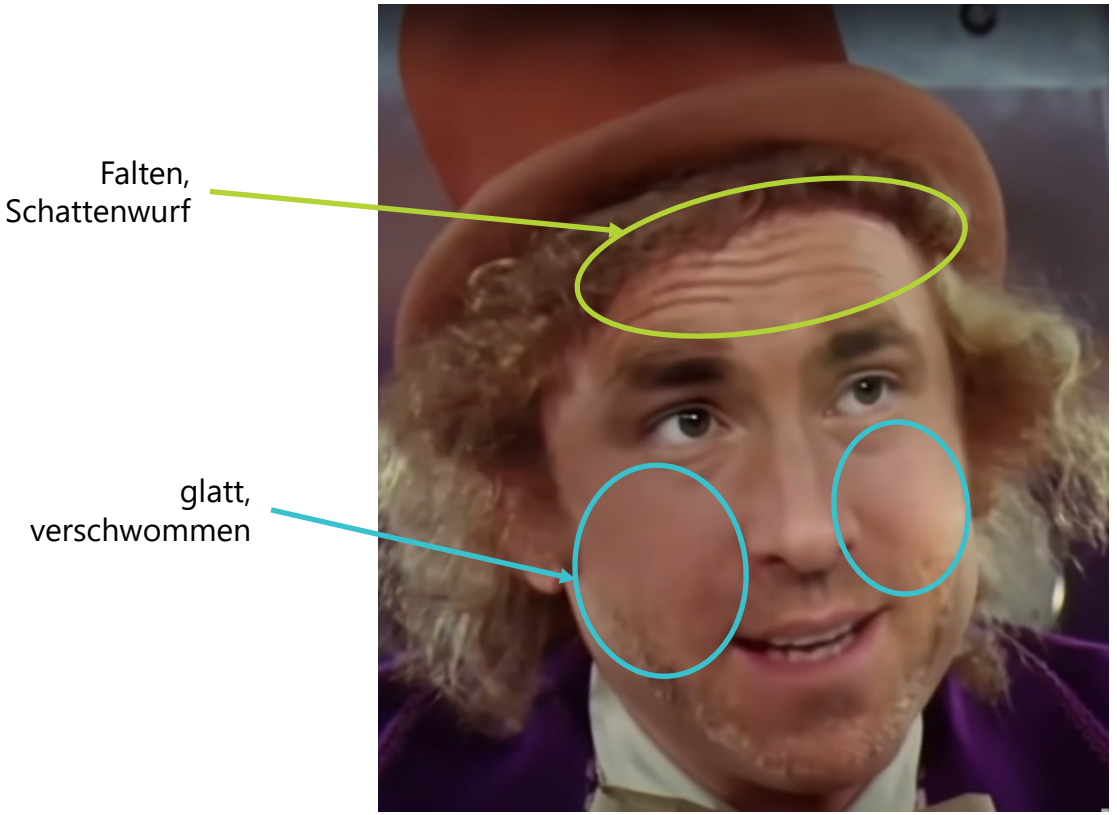
Deepfake Erkennung in Gesichtsregion

Traditionelle Bildforensik

Ansatz:

- Gesichtssegmentierung
- dann automatisierte Analyse bzgl. Textureigenschaften oder Interpolation

Verarbeitungsgeschwindigkeit:
ca. 5 Frames pro Sekunde

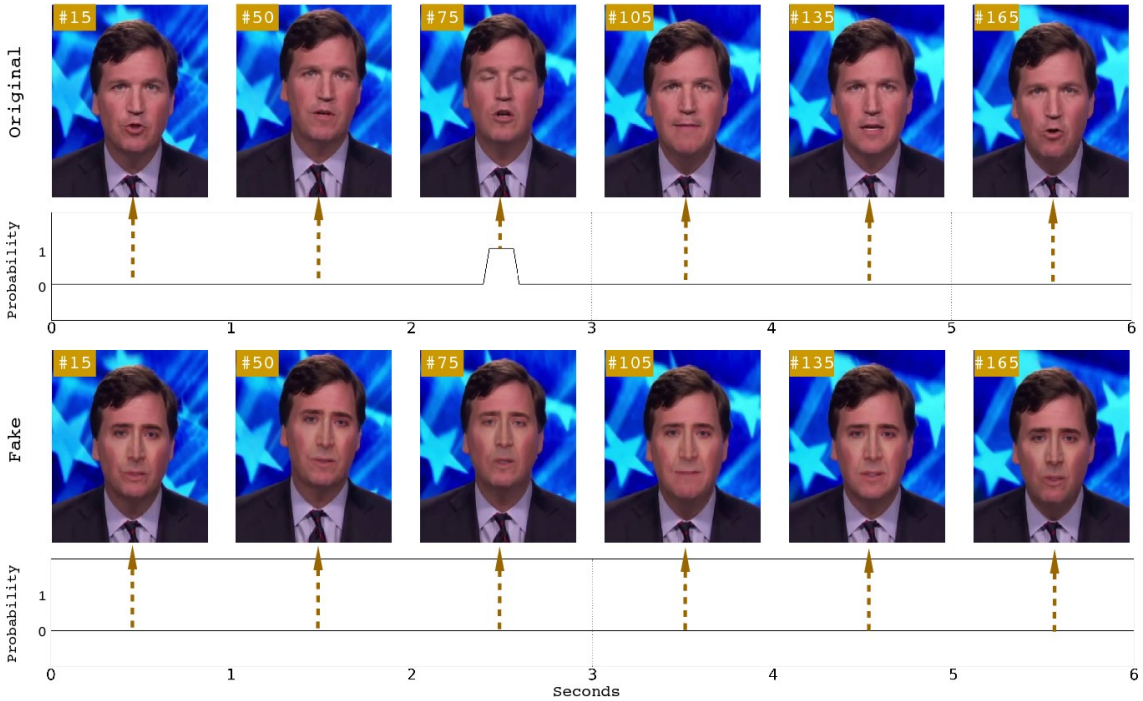


Quellen: Youtube/Watchmojo
https://www.youtube.com/watch?v=Ns0xH29eyjQ (17.01.2024)

Frick, Zmudzinski et al.: Detecting Deepfakes with Haralick's Texture Properties (2021)

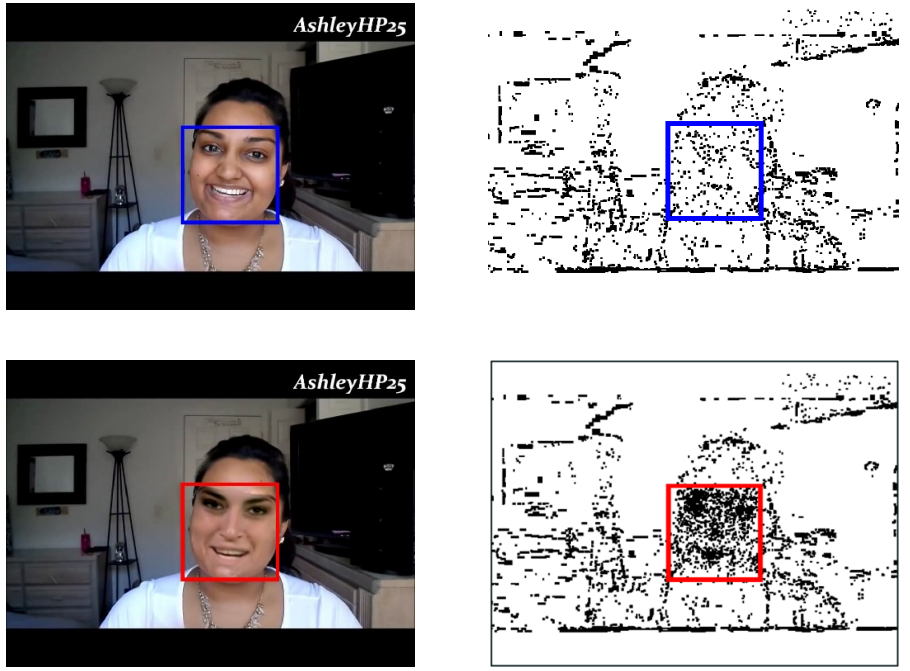
Deepfake Erkennung

Traditionelle Bildforensik



Ansatz: Auswertung des Augenblinzelns

Quelle: Li et al.: In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking (2018)



Ansatz: Auswertung der Kompressionseigenschaften

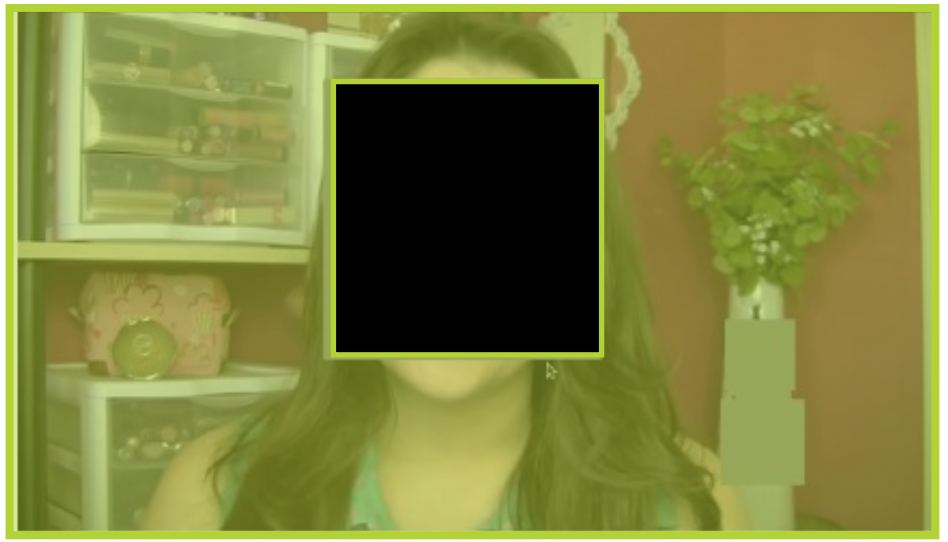
Quellen: FaceForensics++,
Frick, Zmudzinski et al.: Detecting "DeepFakes" in H.264 Video Data Using Compression Ghost Artifacts (2020)

Deepfake Erkennung im Bildhintergrund

Traditionelle Bildforensik

Ansatz: Image-Matching

- Segmentierung Gesicht / Hintergrund
- dann **Bildhintergrund** ggf. bekanntem Bildmaterial zuordnen
  - inverse Bildersuche
  - robustes Video-Hashing



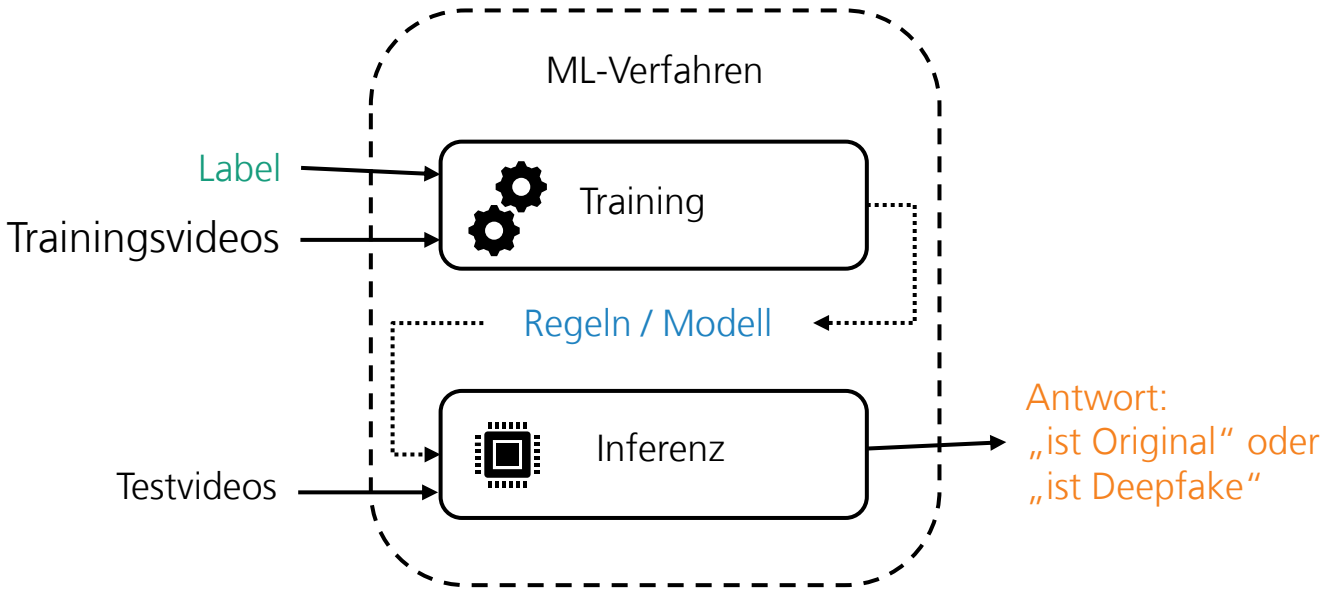
Quellen:
FaceForensics++ Dataset,
Blümer, Steinebach et al.:
Detection of Deepfakes Using
Background-Matching (2023)

Deepfake Erkennung in Gesichtsregion

KI-basierte Erkennung

Ansatz: maschinelles Lernen (ML) zur Erkennung von Deepfake-Angriffen

Beispiel: Antrainieren eines neuronalen Netzes zur Erkennung von Deepfakes (CNN, EfficientNet Architektur)



Videodatensatz:

- 1000 authentische Videos
- 1000 Deepfake-Videos
- Spieldauer ca. 5 Stunden
- Beispiel:

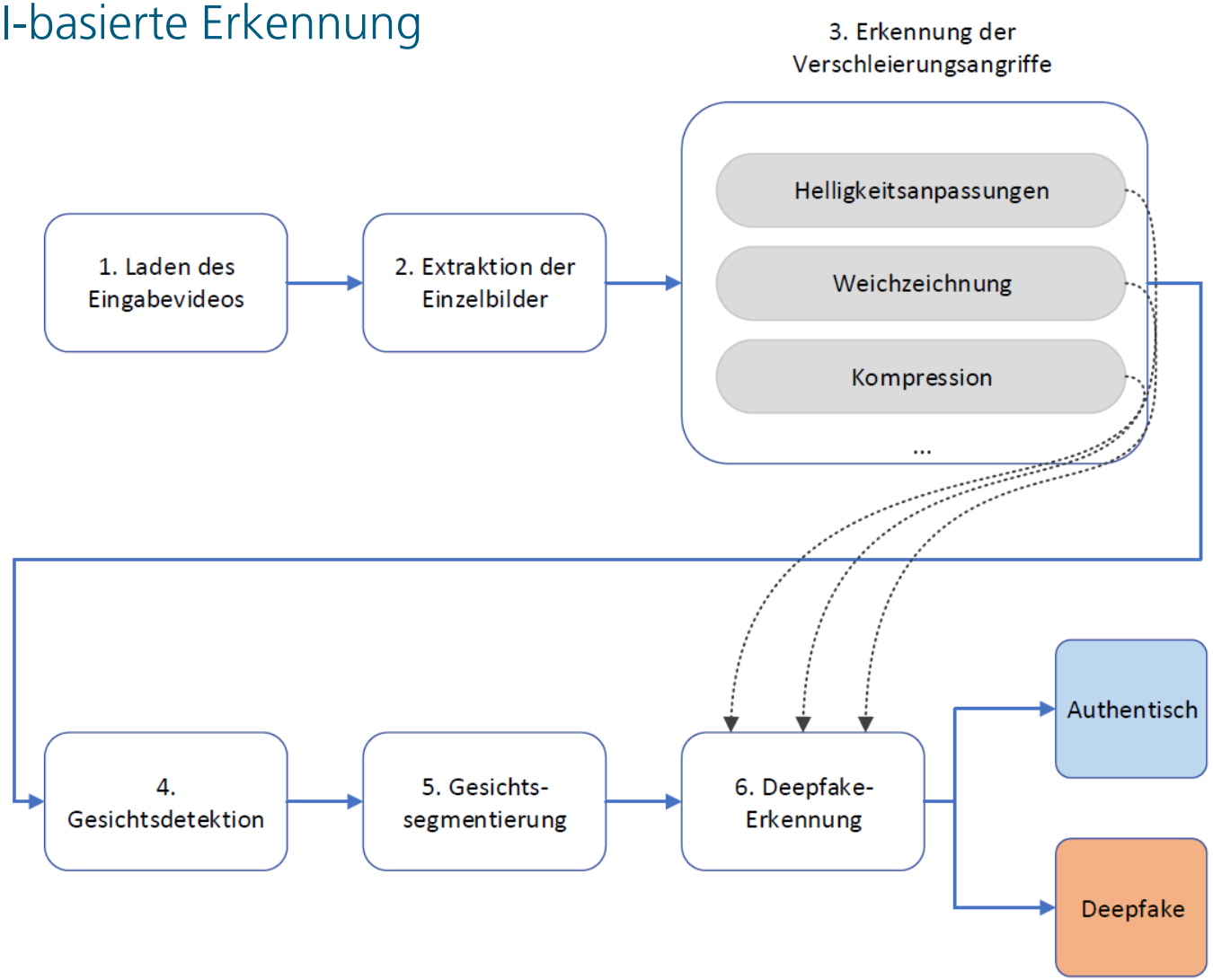
Label „Original“

Label „Deepfake“

Quelle: FaceForensics++ Dataset

Deepfake Erkennung in Gesichtsregion

KI-basierte Erkennung



Ergebnisse

- gute Erkennungsraten im Referenzmodell
- gute Erkennungsraten auch bei schwacher erneuter Videokompression und Bildschärfung
- mit angepasstem, vortrainiertem Deepfake-Classifier: gute Erkennungsraten bei Bildrauschen und Weichzeichnung

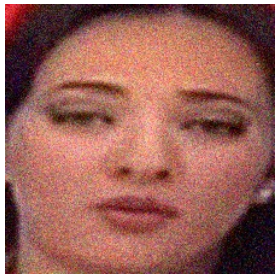
Deepfake Erkennung in Gesichtsregion

KI-basierte Erkennung

Beispiel: Verschleierung durch additives Bildrauschen

1. ML-basierte Erkennung der Verschleierungsart

Ergebnis:

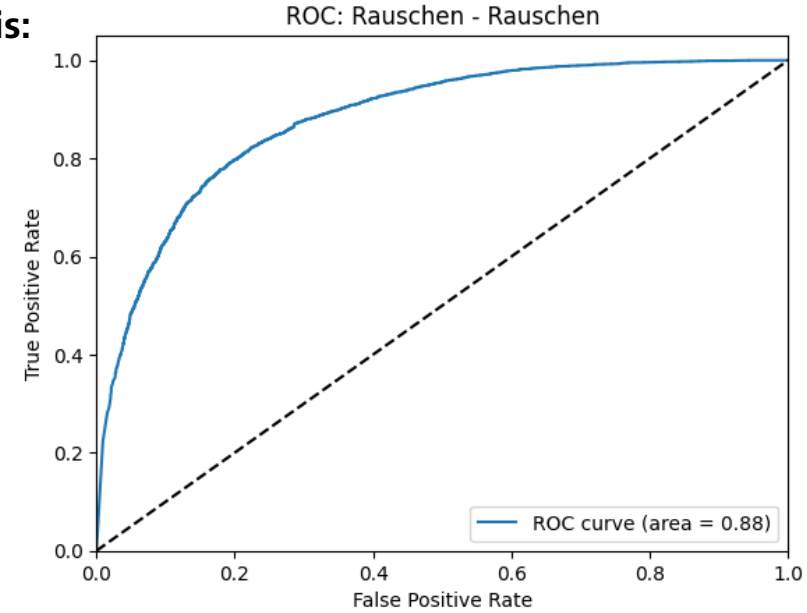


		Tatsächliches Rauschen			
		ohne	**Poisson**	**Speckle**	**Gaussian**
geschätztes Rauschen	**ohne**	**100,0%**	0,0%	0,0%	0,0%
	Poisson	0,0%	**68,3%**	31,7%	0,0%
	Speckle	2,0%	1,7%	**93,1%**	3,2%
	Gaussian	0,0%	0,0%	0,0%	**100,0%**

korrekte Klassifizierungsrate

2. dann speziell vortrainierten Deepfake-Classifier einsetzen

Ergebnis:



→ Ähnlich hohe Trennschärfe wie für das Referenzmodell (ohne Verschleierung)

Zusammenfassung und Ausblick

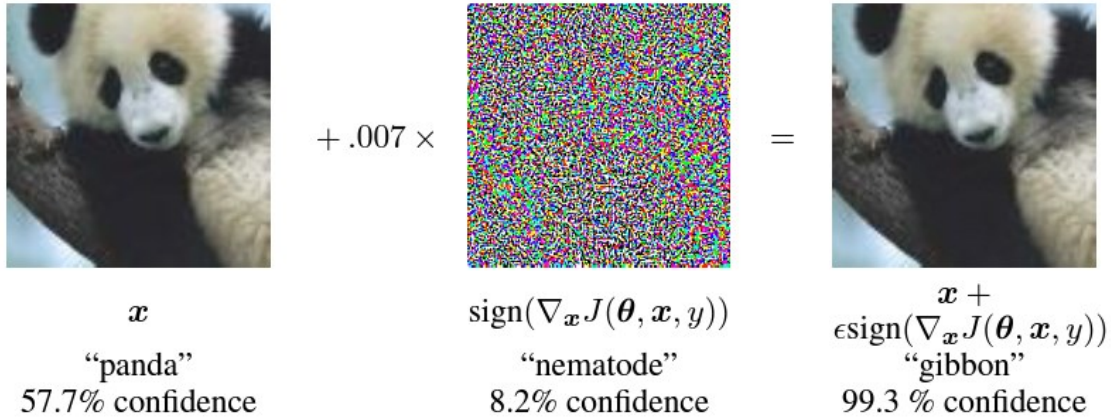
Erkennung von Video-Deepfakes

Stand der Technik, Herausforderungen … und Chancen

- Viele Ansätze für Deepfake-Detektion verfügbar
- Verringerte Erkennungsraten durch starke Bildbearbeitung
- Viele Detektions-Verfahren (derzeit) nicht echtzeitfähig
- Ergebnisse in der Fachliteratur oftmals schlecht zu reproduzieren und auf anderes Videomaterial zu übertragen
- Deepfake-Algorithmen und GPU-Hardware immer leistungsfähiger
- Synchronizität zwischen Bild und Tonspur im Videoclip prüfen
- Adaptive Angriffe, z.B. Adversarial Noise Attacks, berücksichtigen

Unsere Aktivitäten hierzu

- Projekt "SecMedID: Absicherung medialer Identitäten"
  - Face Swapping, auch: Reenactment, Voice Cloning, Text-to-Speech
- Projekt „DREAM: Deepfake REcognition and Artificial Media"
  - auch: Text-to-Image
- Projekt "RoMa: Robustness in Machine Learning"



Beispiel für Adversarial Noise

Quelle: Goodfellow et al.: Explaining and Harnessing Adversarial Examples (2015)

© Fraunhofer

Vielen Dank für Ihre Aufmerksamkeit

Fraunhofer SIT

Fraunhofer SIT

Kontakt

Dr. Sascha Zmudzinski
Abt. Media Security and IT Forensics
Fraunhofer Institut SIT | ATHENE-Center
Sascha.Zmudzinski@sit.fraunhofer.de

Ressourcen:

Webseite: https://www.sit.fraunhofer.de/de/itforensics/
Projekt DREAM: https://revise.athene-center.de/projekte
Projekt RoMa: https://senpai.athene-center.de/projekte#c7153