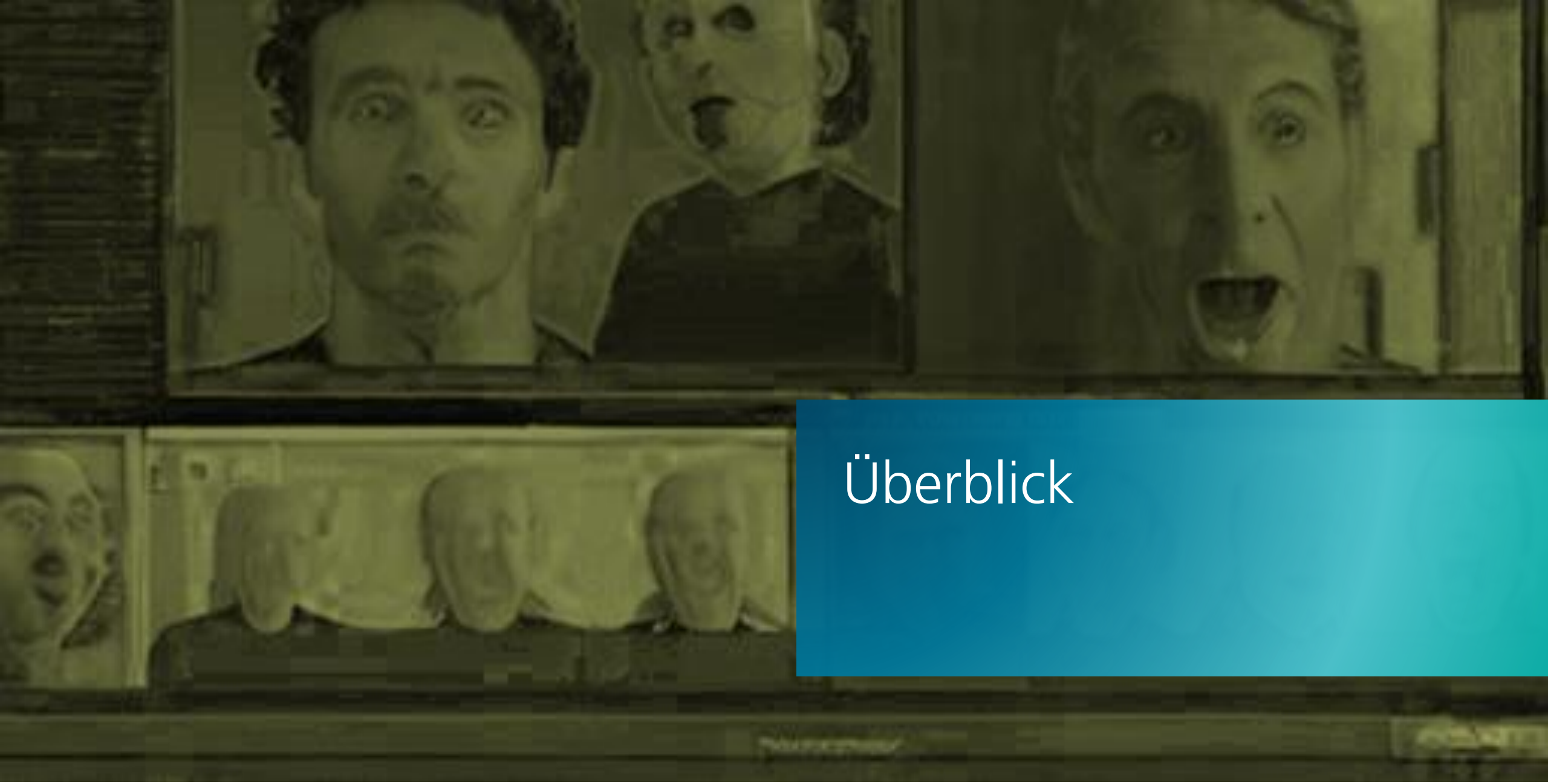
Fraunhofer SIT

Martin Steinebach

Erkennung von Deepfakes und GenAI

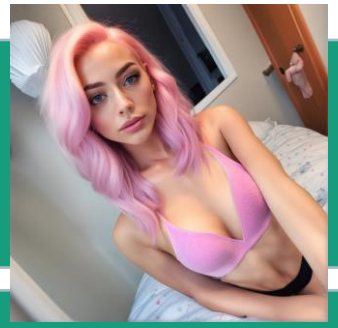
Digitale Beweise III: Deep Fake



Überblick

KI-basierte Manipulationen

https://www.reddit.com/r/StableDiffusion/comments/14svyoa/sdxl09_can_it_do_nsfw/



Image

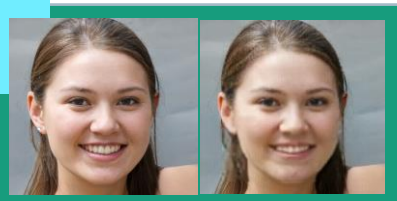
- Neural Editing
- Text-to-Image
- Image-to-Image

Image-to-Video

Video

- Deepfakes
- Text-to-Video
- Video-to-Video

Lipsync



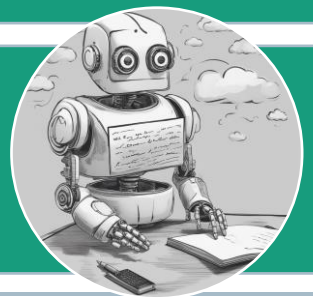
Overdub

Audio

- Text-to-Speech / Text-to-Sound
- Speech-to-Speech
- Voice Cloning

Text

- LLMS



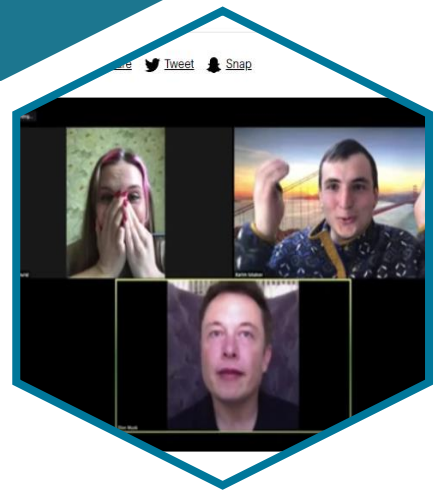
Warum beschäftigten wir uns damit?

GenCSAM

Desinformation

Mobbing

Betrug



A) Mage.space / stable diffusion
B) https://aimojo.pro/top-10-free-undress-ai-tools-exploring-safe-practices-for-image-manipulation
C) X/Eliot Higgins
https://www.20min.ch/story/donald-trump-wurde-mithilfe-kuenstlicher-intelligenz-ki-verhaftet-fake-bilder-auf-twitter-viral-530578576348
D) https://www.vice.com/en/article/g5xagy/this-open-source-program-deepfakes-you-during-zoom-meetings-in-real-time

Sicherheitskonzepte

Aktiv

Passiv

- Watermarking
- Filter

- Watermarking
- Filter
- Listen

- Forensik

Sicherheitskonzepte – Verbreitung

Forensik

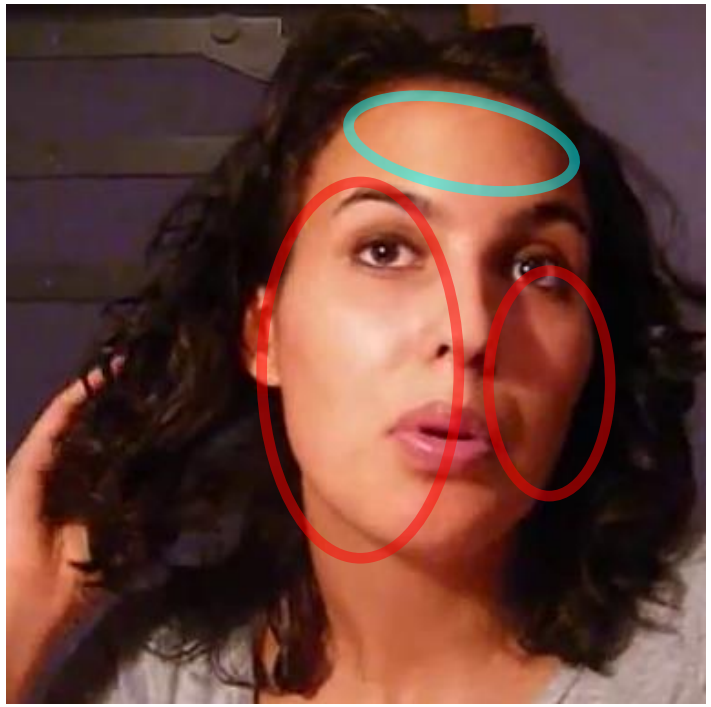
- Erkennung von Erstellungsspuren
  - Upscaling-Muster
  - Hochfrequente Audiomuster
- Erkennung von typischen Stilen
  - NLP ChatGPT-Erkennung (plain vanilla)
- Erkennung von Erstellungsfehlern
  - Unstimmigkeiten bei der Hautfarbe (Deepfakes)
  - Unstimmigkeiten bei der Schärfe (Text zu Bild)

Herausforderungen: Fehlerquoten (Fehlalarme)

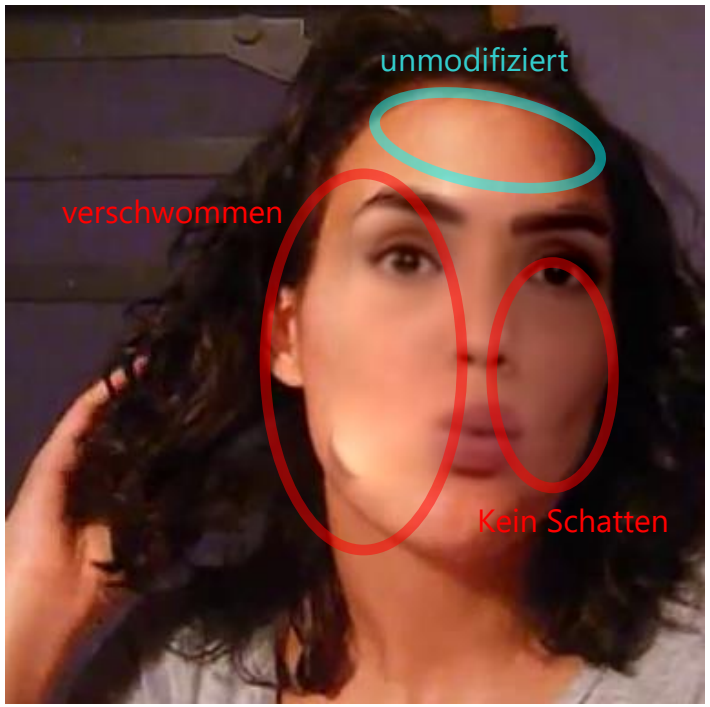
Technische Verfahren

Deepfake Erkennung in Gesichtsregion

Original Video Frame



Deepfaked Video Frame

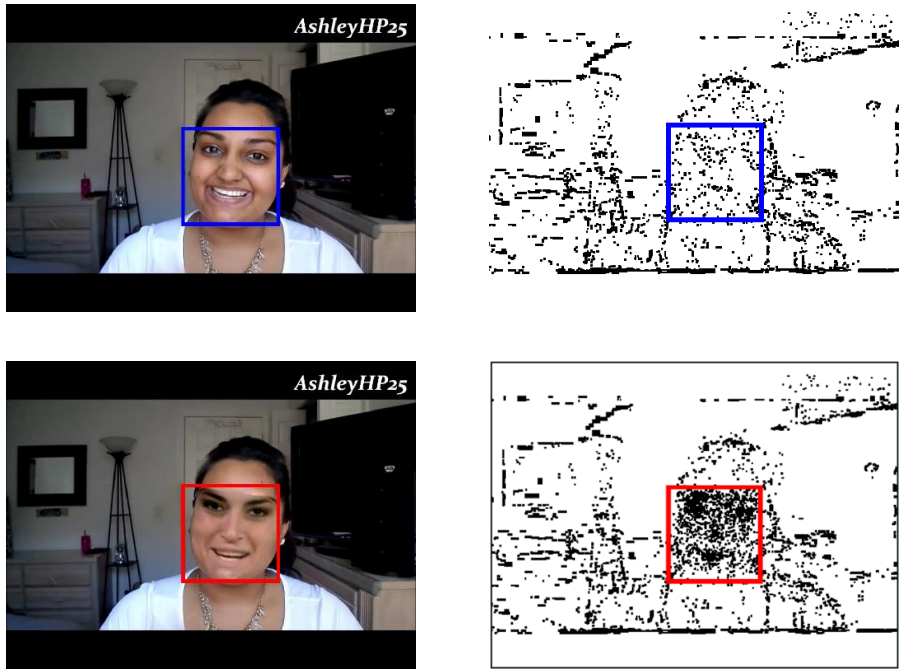


Difference Mask

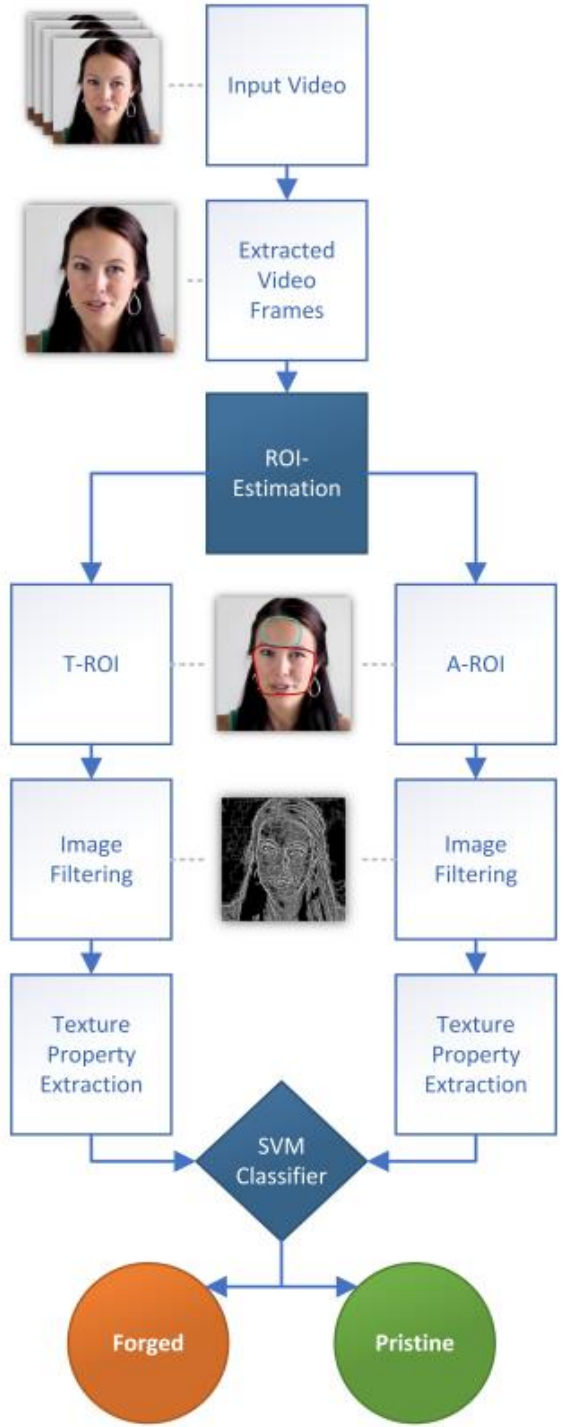
- Vergleich von Eigenschaften in ähnlichen (benachbarten) Regionen
- Analyse von Textureigenschaften
- 3 bis 5 Frames pro Sekunden können auf Spiele-PC so analysiert werden

Source: FaceForensics++ Dataset

Deepfake-Erkennung

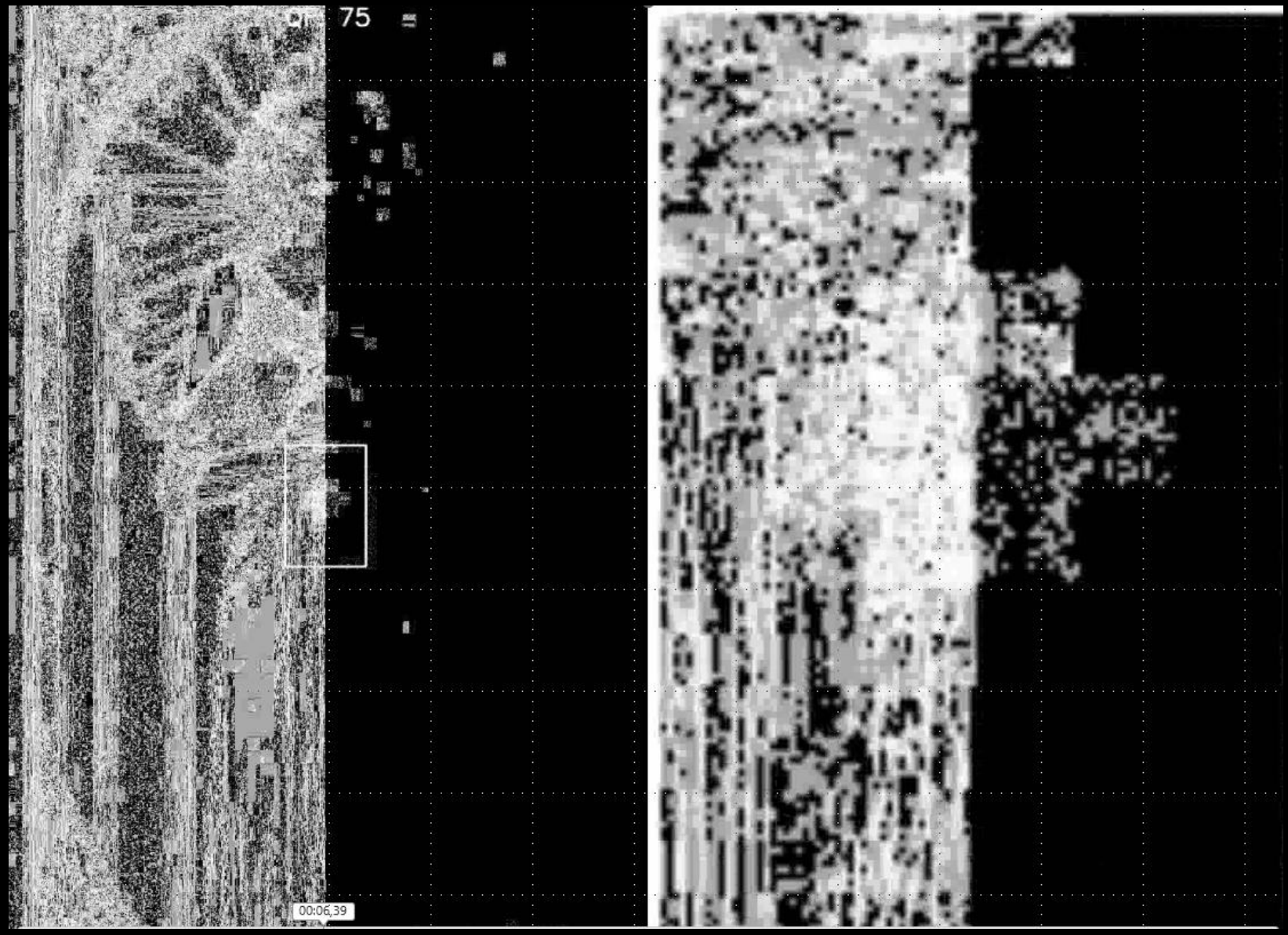


Auswertung der Kompressionseigenschaften
(blau = authentisch, rot = Deepfake)
Quellen: FaceForensics++ Dataset; Detecting "DeepFakes" in H.264 Video Data Using Compression Ghost Artifacts, Frick et al.



Beispiel

L:80 R:75



Erkennen von Text-to-Image

- Automatisiert durch forensische Methoden
  - Basierend auf dem Erzeugungsprozess
- Automatisiert durch maschinelles Lernen
- Digitale Wasserzeichen
  - Beim Training
  - Nach dem Erzeugen
- Manuelles Suchen nach Fehlern

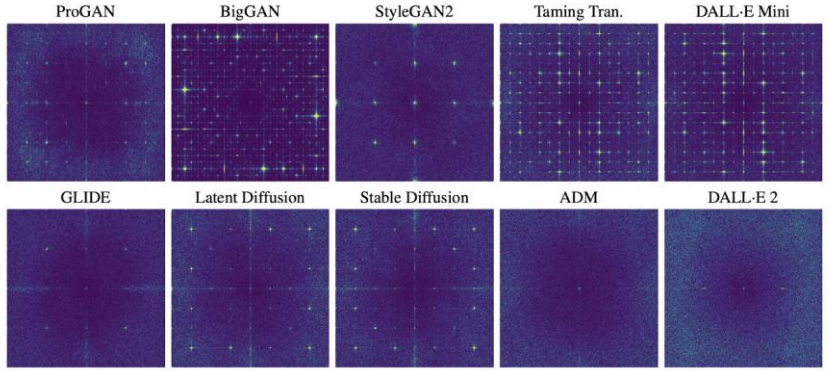


Fig. 2: Fourier transform (amplitude) of the artificial fingerprint estimated from 1000 image residuals. Top row: from left to right ProGAN [20], BigGan [21], StyleGAN2 [22], Taming Transformers [23], DALL·E Mini [24]. Bottom row: GLIDE [5], Latent Diffusion [25], Stable Diffusion [4], ADM [26], DALL·E 2 [3]

ON THE DETECTION OF SYNTHETIC IMAGES GENERATED BY DIFFUSION MODELS
https://arxiv.org/pdf/2211.00680.pdf

Optische Merkmale

Erkennen von Text-to-Image

Beispiele für manuelle Erkennung

Auffällige Augenform bzw. Fehler in der Darstellung von Augen

Erkennen von Text-to-Image

Beispiele für manuelle Erkennung

Anzahl von Fingern nicht korrekt

https://www.reddit.com/r/StableDiffusion/comments/x9vjka/hand/

Erkennen von Text-to-Image

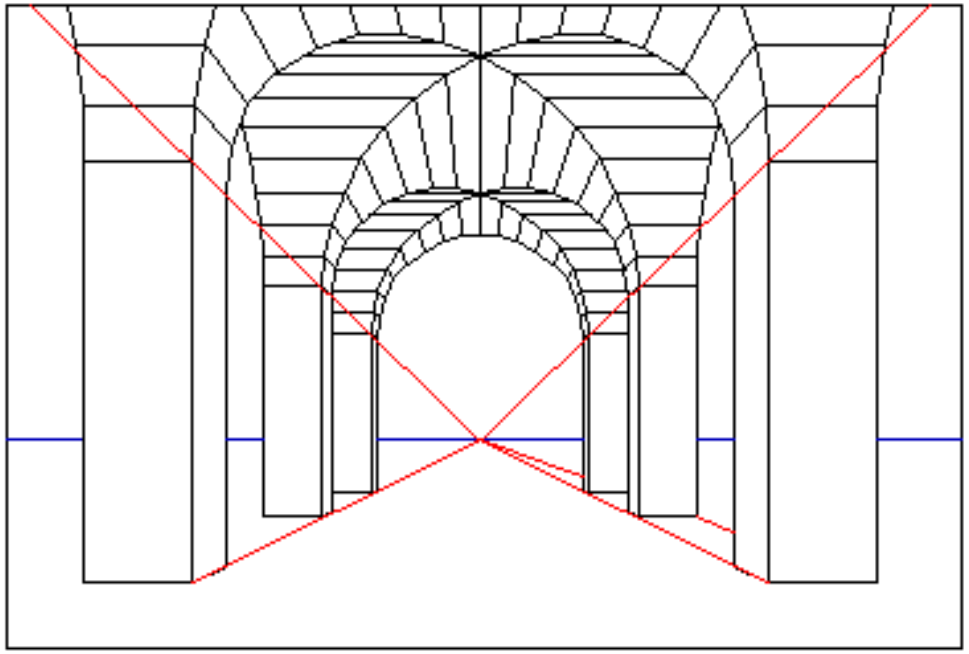
Beispiele für manuelle Erkennung

Fehler in Objektübergängen

Erkennen von Text-to-Image

Beispiele für manuelle Erkennung

Fehler in Fluchtpunkten



https://commons.wikimedia.org/wiki/File:Zentralperspektive.png

Erkennen von Text-to-Image

Erkennen von Text-to-Image

Spiegelung nicht korrekt

Erkennen von Text-to-Image

Beispiele für manuelle Erkennung

Fehler in Tiefenschärfe

Liar’s Dividend

Abstreitbarkeit ist ein wichtiges Risiko von Deepfakes

Zusammenfassung

- Deepfakes und GenAI bezeichnen eine ganze Familie von Manipulationsverfahren
  - für Bild, Audio, Video und Text
  - Erstellen neuer oder Ändern bekannter Inhalte
- Technische Erkennungsverfahren basieren auf dem Herausarbeiten von Spuren
  - Interpretation der Spuren erfordert technische Kenntnisse
  - Fehlerraten sind abhängig von Verfahren und Szenario
- Manuelle Erkennung basiert auf Fehlern im erzeugten Inhalt
  - Belastbar wenn vorhanden
  - Fehler werden nicht zwingend erzeugt, daher keine Beweiskraft bei fehlerfreien Inhalten

Vielen Dank

Kontakt

Prof. Dr. Martin Steinebach
Head of Media Security and IT Forensics
Tel. +49 6151 869-349
martin.steinebach@sit.fraunhofer.de

https://www.sit.fraunhofer.de

Fraunhofer SIT