Friedrich-Alexander-Universität
Technische Fakultät

Cybercrime and Forensic Computing
Research Training Group 2475

FAU

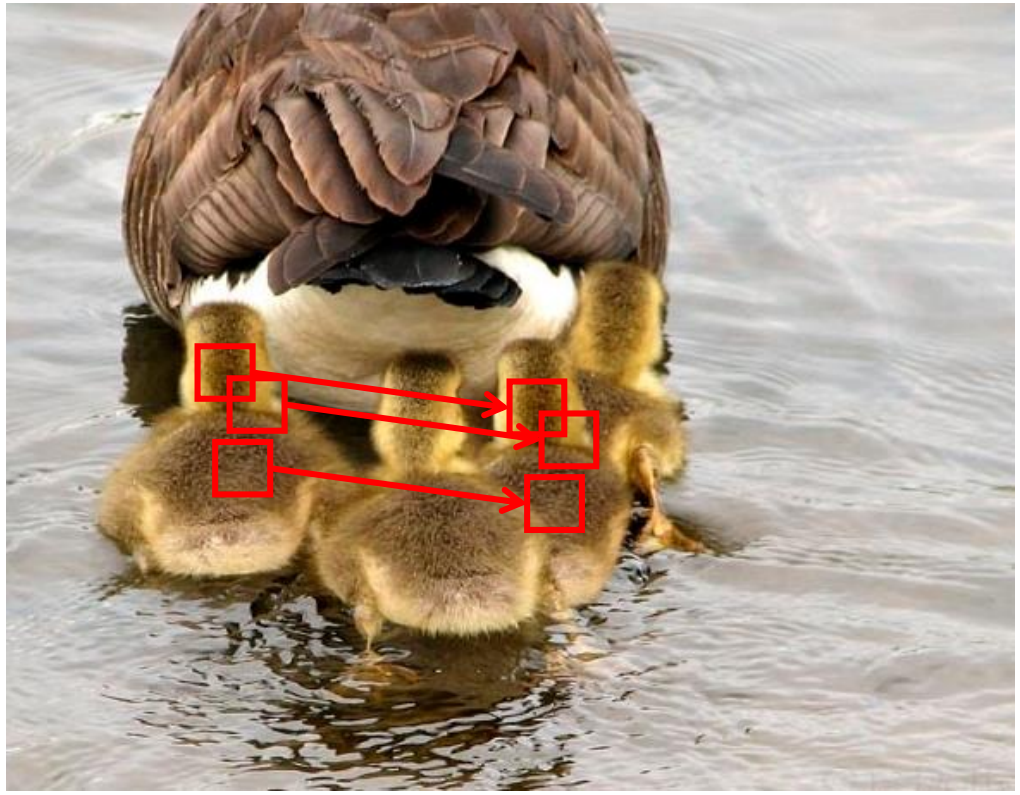
Generative KI: Erzeugung und Erkennung computergenerierter Bilder

EDV-Gerichtstag
Saarbrücken
13.9.2023

PD Dr.-Ing. Christian Riess
Lehrstuhl für IT-Sicherheitsinfrastrukturen
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)

Bildmanipulationen und Manipulationserkennung “früher”

- Während meiner Promotion (2007-12) war Bildbearbeitung eine manuelle Kunst
- Z.B.: meine wichtigste Arbeit untersuchte Erkenner für kopierte Regionen



Automatisierte Bildverarbeitung klopft an das Tor

- Manuelle Bildmanipulationen (“Shallow Fakes”, “Cheap Fakes”) sind für journalistische Faktenprüfer nach wie vor am häufigsten und relevantesten

Julia Bayer, Deutsche Welle

Patrick Gensing, NDR

- Aber:

  Seit 2021 kündigen Text-zu-Bild-Generatoren wie Dall-E, midjourney, stable diffusion einen wesentlich höheren Grad von Automatisierung an

Angeblich betrunkene N. Pelosi: Verlangsamtes Abspieltempo

Wie benutzt man Text-zu-Bild-Generatoren?

- Textuelle Beschreibung eingeben, was erzeugt werden soll:

  `> police officers having donuts in their office at dusk`

- Kurz warten…
- Voila:

- Ich finde es interessant, dass auch wenig hilfreiche Beschreibungen wie „in their office at dusk“ oft ganz OK gelöst werden
- Dennoch sind die Modelle noch im Frühstadium: Fehler in Frisur und den Händen

Wie funktionieren Text-zu-Bild-Generatoren intern?

- Ein neuronales Netz sucht eine kompakte interne Darstellung für einen Textabschnitt (Latent Space), um daraus die nächsten Sätze vorherzusagen:
- **“hier geht um Hautcremes, die nächste Sätze könnten z.B. Nebenwirkungen besprechen”**
- Ein zweites neuronales Netz sucht Bild-Text-Paare, und sucht für jedes Bild **die selbe interne Darstellung** wie der dazugehörige Text

“hier geht es um Hänsel und Gretel, das Bild könnte ein ikonisches Element dieses Märchens zeigen, z.B. Hänsel, Gretel, die Hexe, oder das Hexenhaus”

- Beide Modelle werden dann zusammengeschlossen: Der Benutzer gibt dem ersten Netz Text, und das zweite Netz erzeugt daraus ein passendes Bild

KI auf Steroiden

- Gute Beschreibungen für ein Bild zu finden ist oft nicht so einfach
- Findige technik-affine Nutzer können diese Aufgabe aber an eine KI auslagern…

„female influencer from the 1930s“

Unterscheidung von Fotografien von generierten Bildern

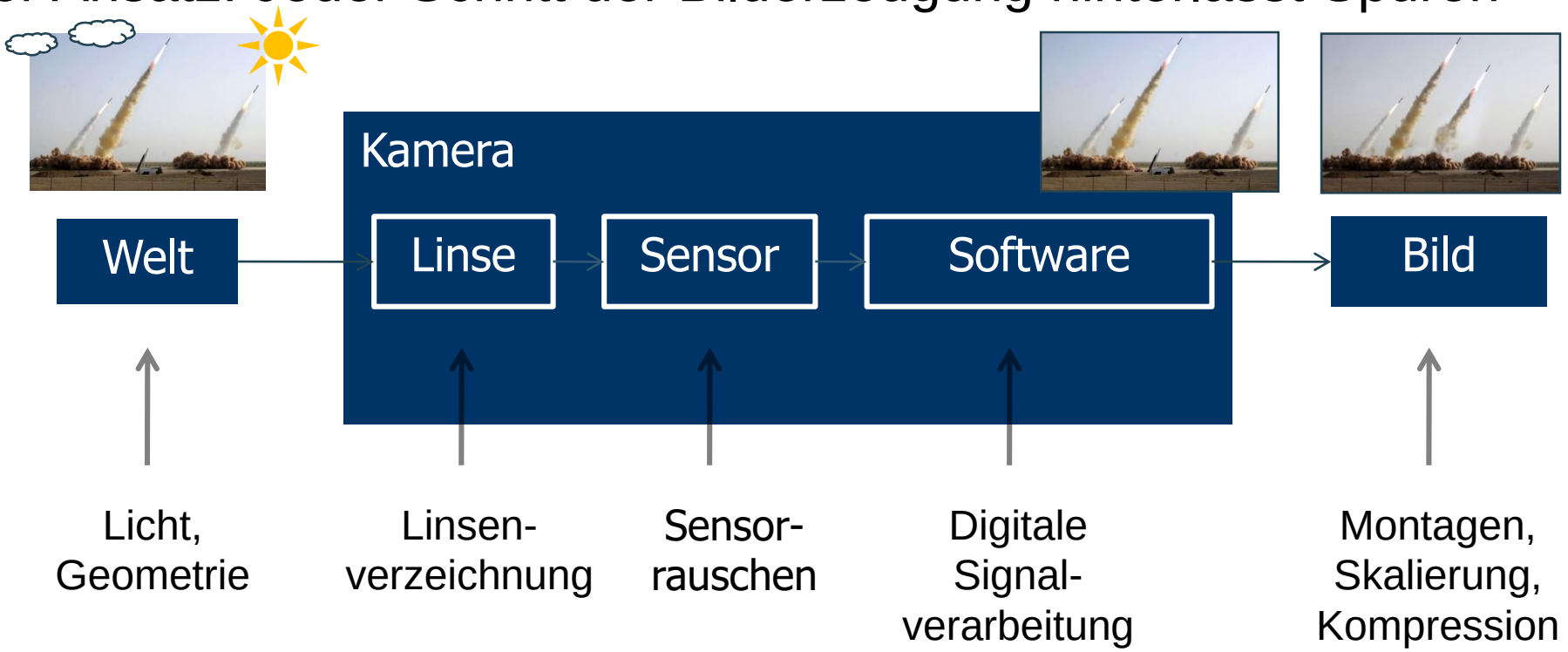
- Aktiv:
  Zusätzliches digitales Wasserzeichen einfügen.
  Vorteil: unsichtbare, robuste Markierungen für KI-Bilder
  Nachteil: Der Anbieter des Netzes muss dies unterstützen

- Passiv:
  Suche nach forensischen Spuren von generierten Bildern
  Vorteil: Unabhängig vom Anbieter
  Nachteil: Erfolg ist situationsabhängig

- Im Folgenden werden passive Methoden weiter besprochen

Bildforensische Analyse

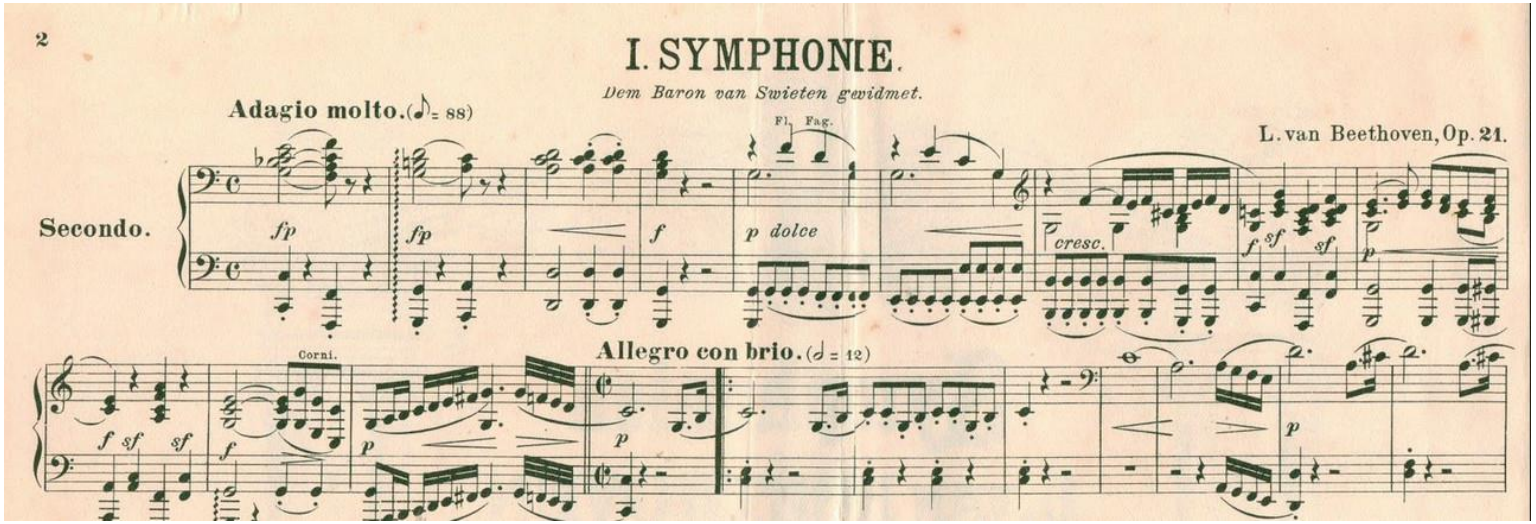
- Analytischer Ansatz: Jeder Schritt der Bilderzeugung hinterlässt Spuren



- KI-generierten Bildern fehlen die ersten Schritte
- Die Spuren werden teils durch Netzwerkspuren ersetzt, teils (planlos) nachgeahmt

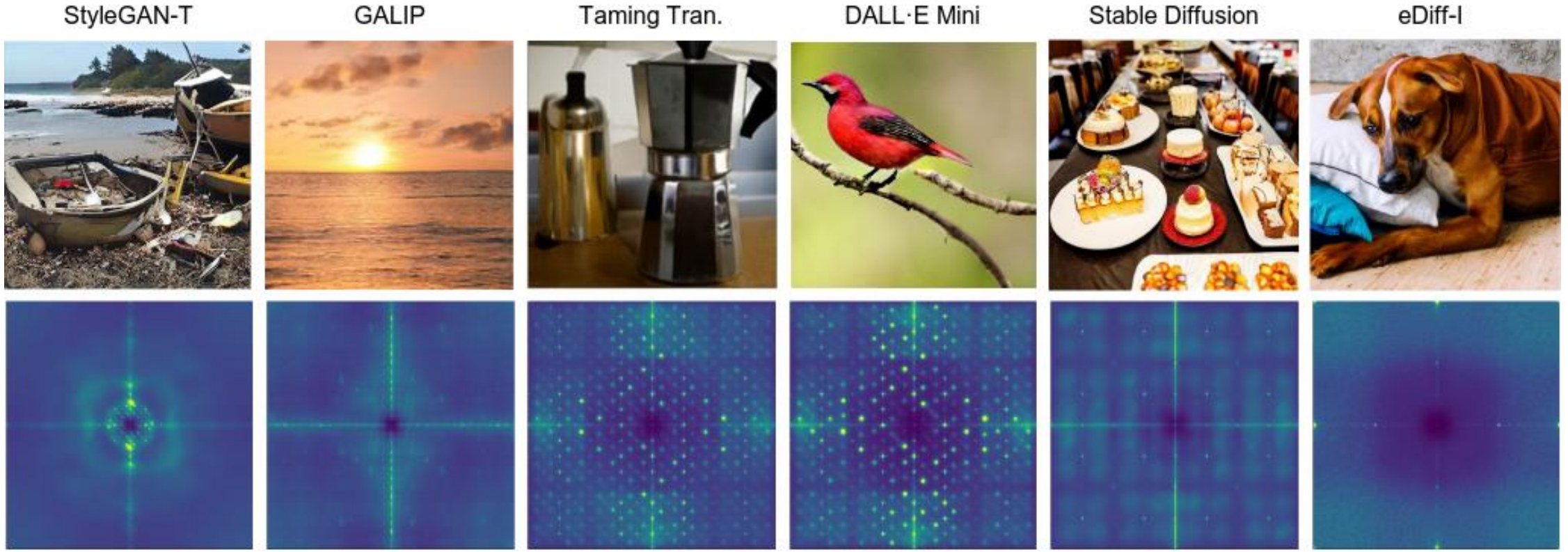
Beispiel: Spuren im Frequenzraum

- In Naturwissenschaften und Technik ist der Frequenzraum eine beliebte alternative Darstellung der Daten
- Statt Pixel für Pixel, wird das Bild als Summe von Schwingungen repräsentiert
- Der Inhalt ist in beiden Darstellungen äquivalent.
  Stellen Sie sich ein Musikstück einmal als Notensatz und einmal als Tonspur vor.



Generierte Bilder erzeugen Spuren im Frequenzraum

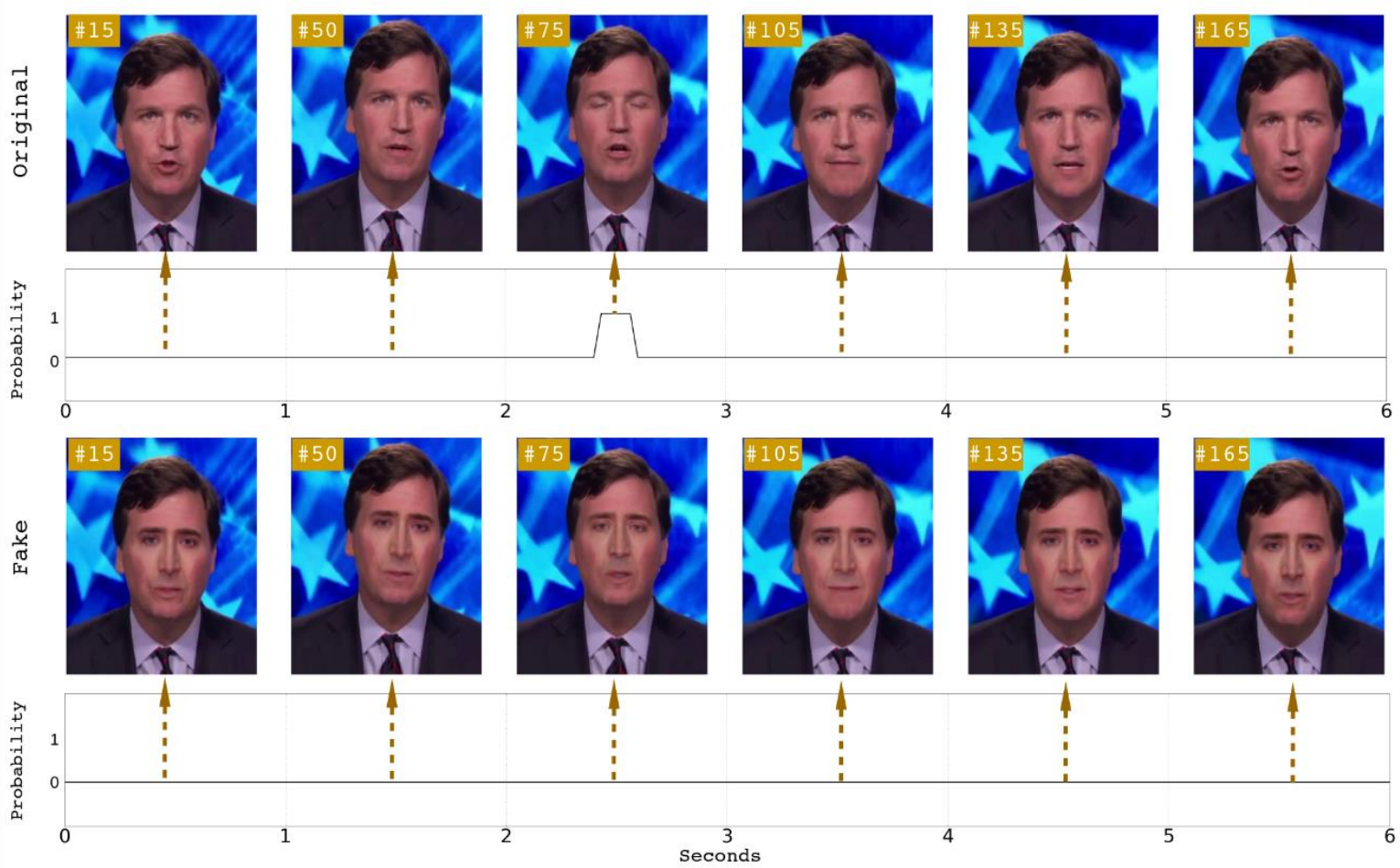
- Verschiedene Generatoren und deren Frequenzraum-Spuren:



- Diese Spuren hängen mit der internen Bilderzeugung zusammen

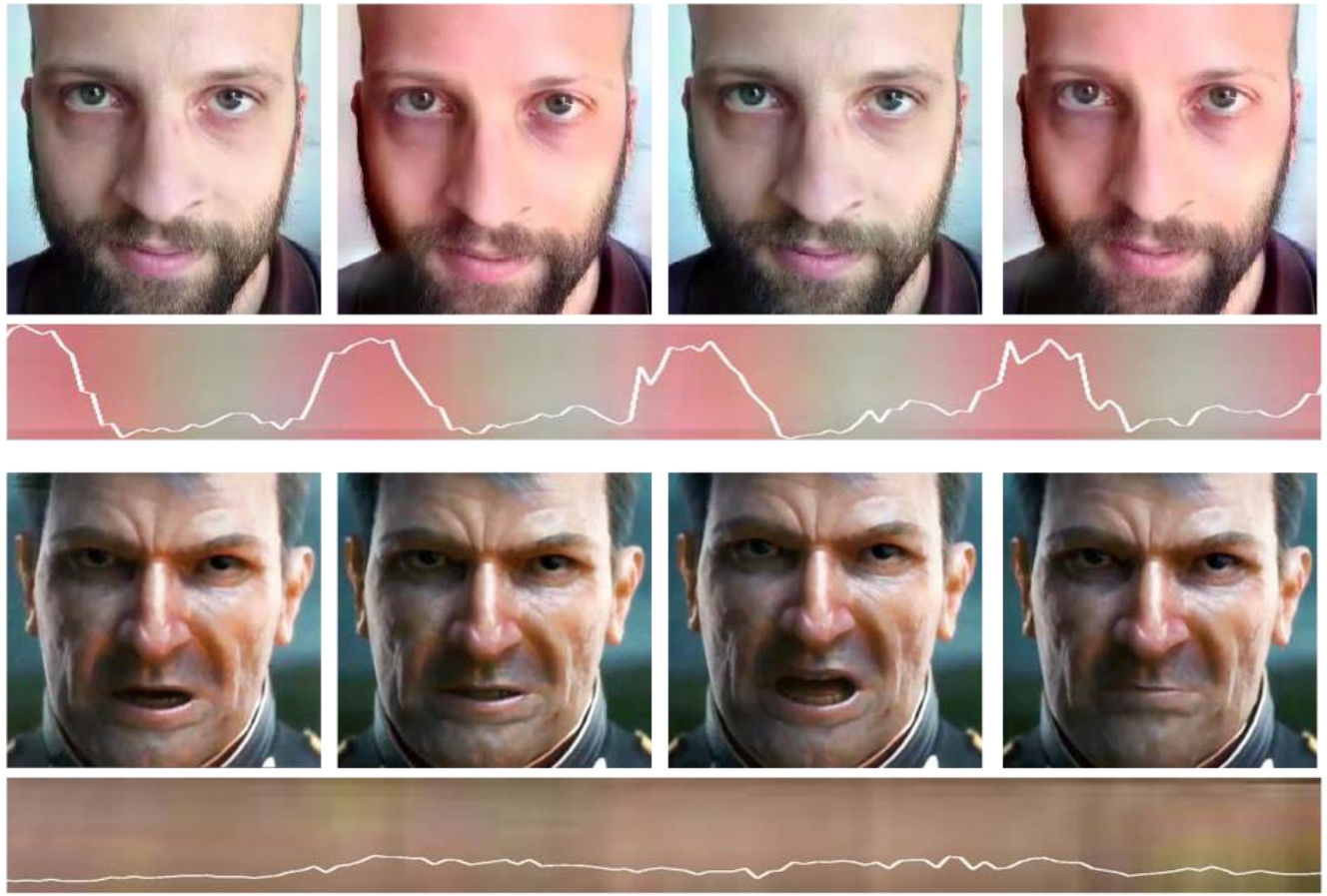
Bild aus: Corvi et al.: „Intruiging Properties of Synthetic Images: From Generative Adversarial Networks to Diffusion Models, CVPRW 2023.

Beispiel für physiologische Spuren in Videos: Zwinkern



Li *et al.* „In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking“, WIFS, 2018.

Beispiel für physiologische Spuren in Videos: Puls



Conotter *et al.* „Physiologically-based Detection of Computer Generated Faces in Video“, ICIP, 2014.

Zusammenfassung und Einordnung

- KI-Bildgeneratoren sind nicht perfekt, zeigen aber enormes Potential, z.B. durch KI, die KI steuert
- Aktive Schutzmaßnahmen durch Wasserzeichen können für populäre Generatoren helfen, aber werden wohl nie von allen umgesetzt werden
- Passive forensische Verifikation ist weniger zuverlässig (z.B. verschwinden Frequenzspuren unter starker Kompression), aber sie ist herstellerunabhängig anwendbar

Vielen Dank!