DEEP FAKE - VISUELLE MANIPULATION MEDIALER INHALTE

IKZ IM BEREICH DER CYBERSICHERHEIT

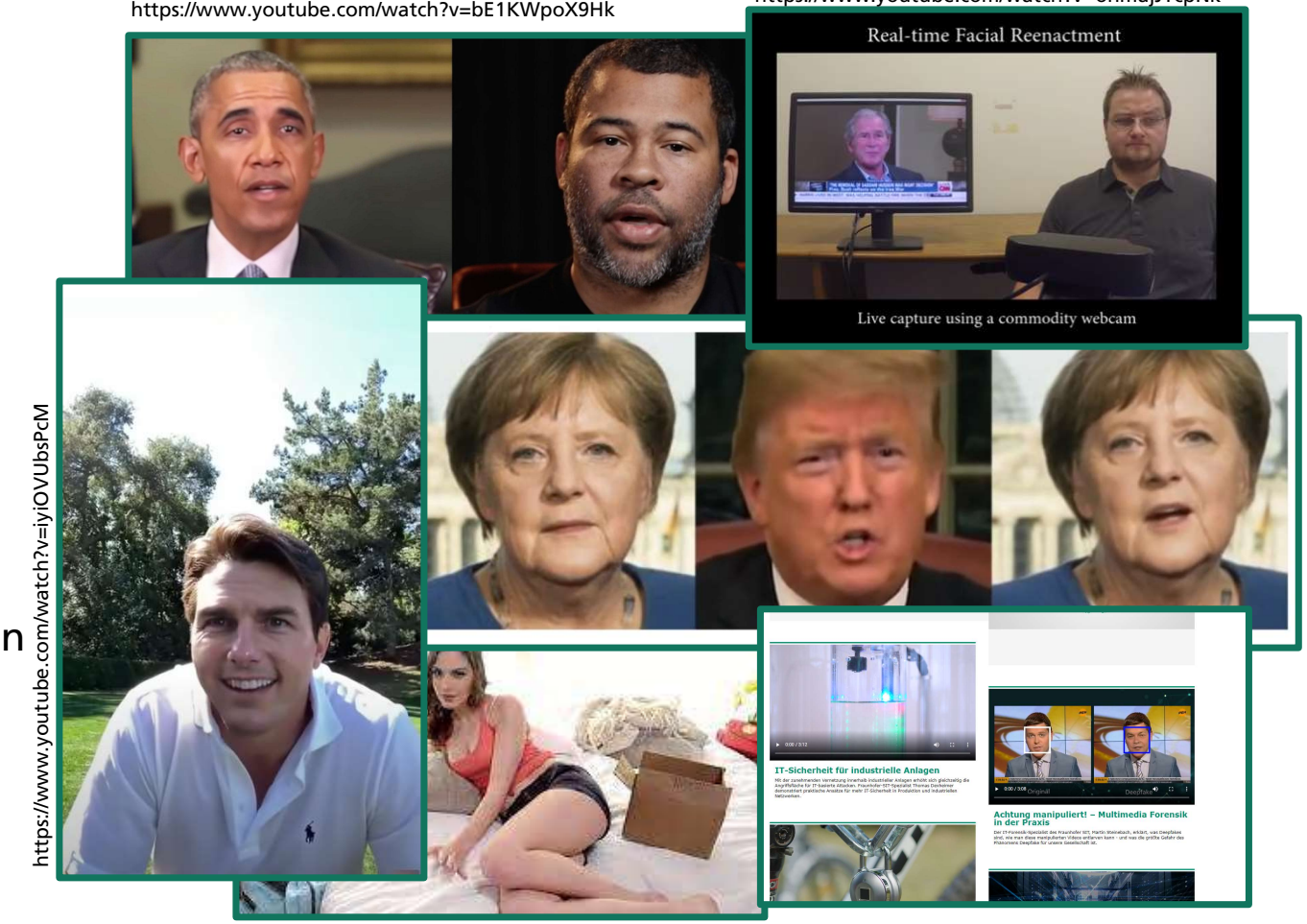
Martin Steinebach

Motivation

- Einsatzfelder von Deepfakes
  - Mobbing
  - Betrug
  - Desinformation
  - Entertainment
  - Optimierung von Kommunikation

https://www.youtube.com/watch?v=bE1KWpoX9Hk

https://www.youtube.com/watch?v=ohmajJTcpNk

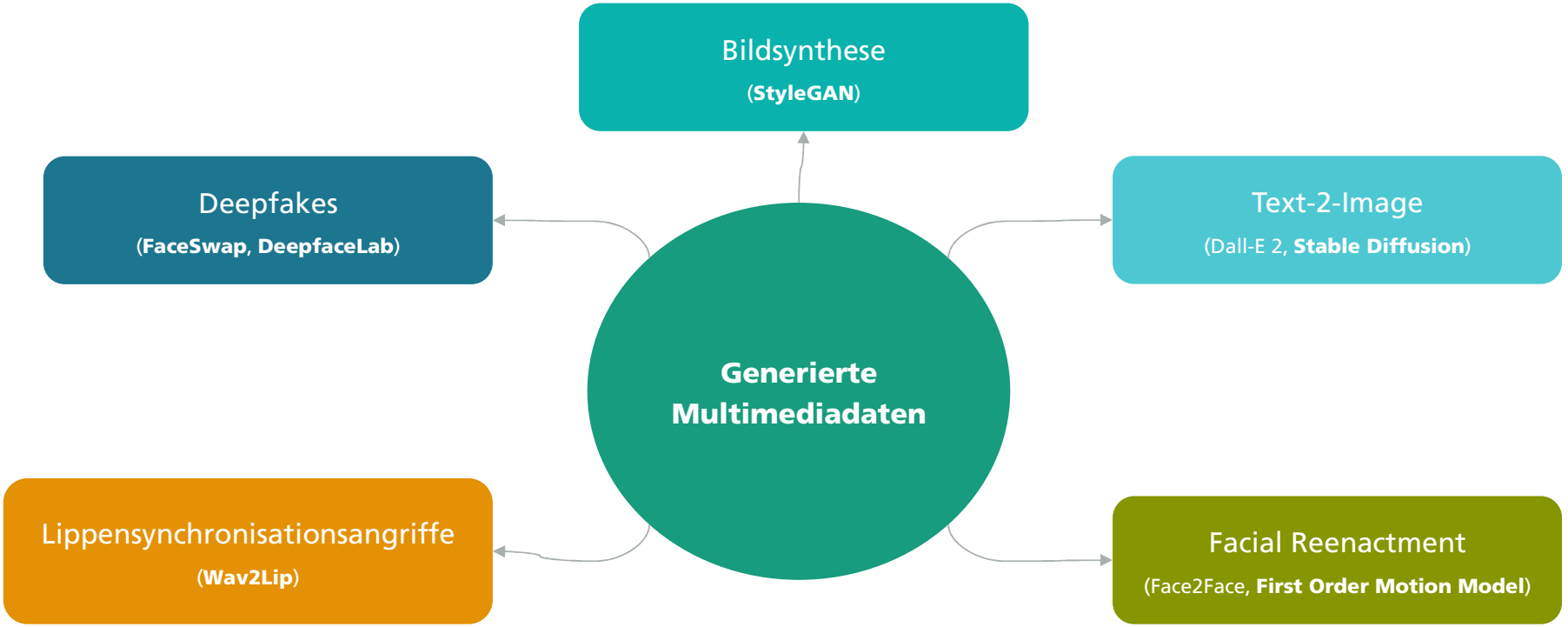


https://www.youtube.com/watch?v=iyiOVUbsPcM

https://innovationslounge.sit.fraunhofer.de

Überblick

- Deepfakes als übergeordneter Begriff für verschiedene Angriffsmethoden



Aktuelle Entwicklung: Echtzeitfähigkeit

- DeepFaceLive
  - Echtzeit Deefakes
  - Benötigt einen normalen Spiele-PC
  - Training offline, Ersetzung in Echtzeit

https://mixed.de/deepfacelive-bringt-deepfakes-in-live-video-streams/

Werkzeug

- Training erfordert wenige Minuten Videomaterial der Zielperson
- Dauer grob 1 Tag auf einem Computer mit guter GPU
- Training kann optimiert werden
  - Auch Quellperson kann trainiert werden

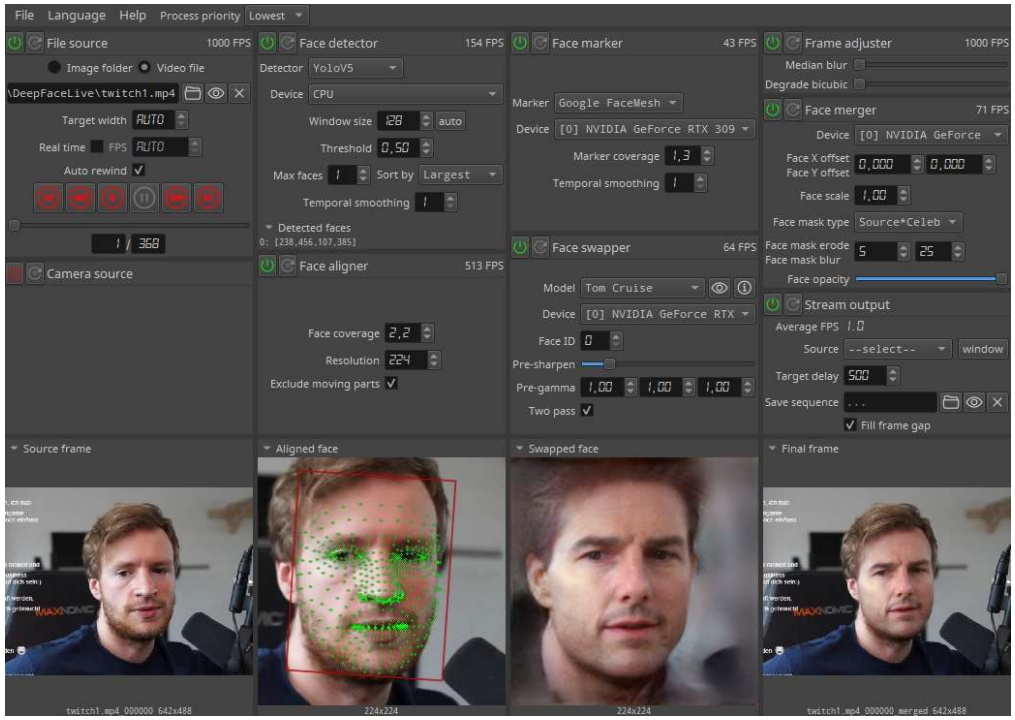
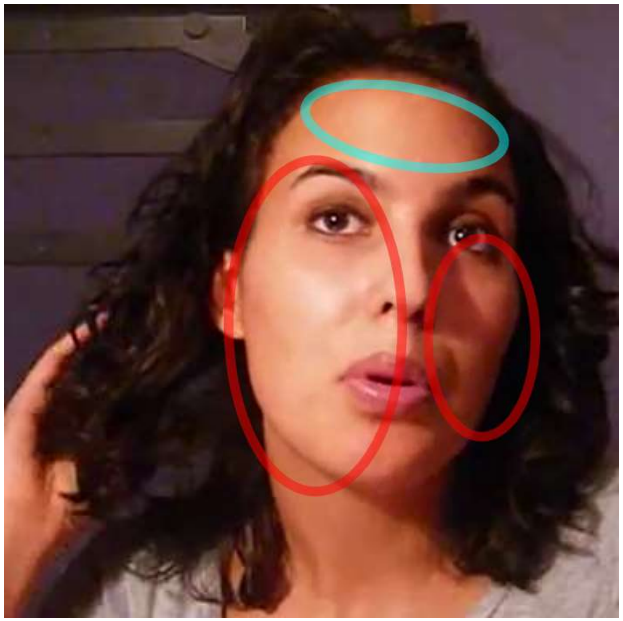


Abb.: Oberfläche von DeepfaceLive (https://github.com/iperov/DeepFaceLive)

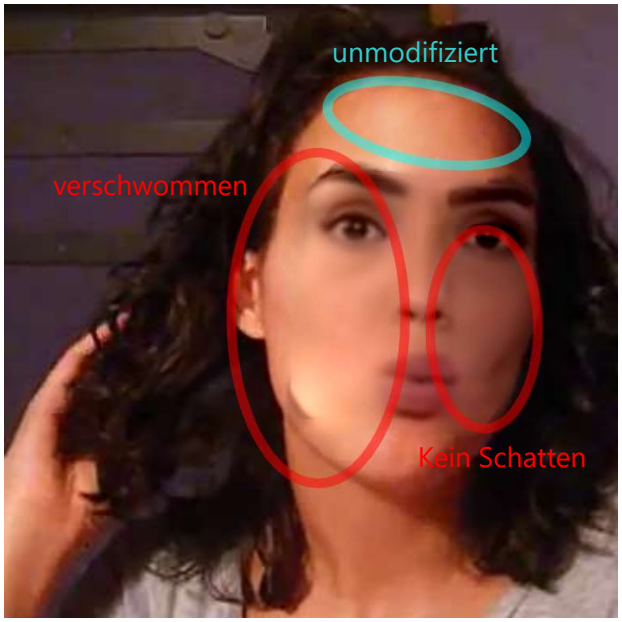

DeepFaceLive output
DeepFaceLive
File Language Help Process priority Normal
File source
Face detector
Face marker
Camera source
Face aligner
Face animator
Source frame
aligned face
DELL

Deepfake Erkennung in Gesichtsregion

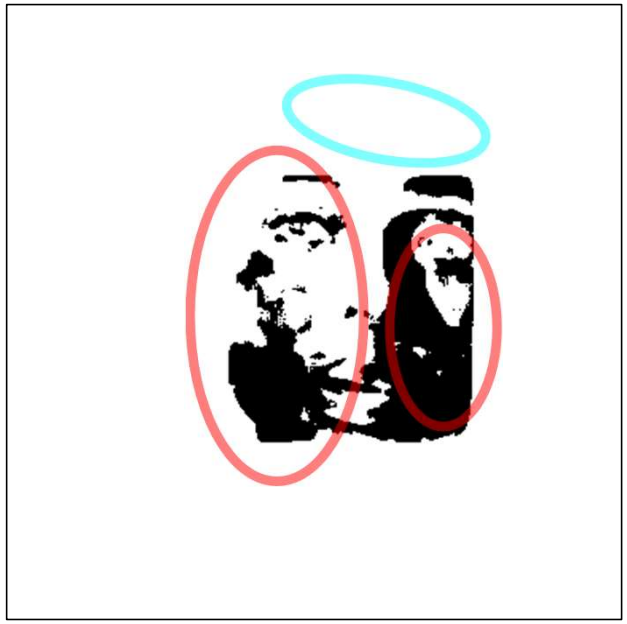
Original Video Frame



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

ATHENE
National Research Center for Applied Cybersecurity

Fraunhofer SIT

Herausforderung Videokonferenzen

- Videokonferenz-Tools bieten Verschönerungsfilter
  - Im Grunde ein Tiefpassfilter
  - Fortgeschrittene Tools unterscheiden zwischen Gesicht und Hintergrund
- Konsequenzen
  - Aufpolierte authentische Videos können als Fälschungen erscheinen
  - Deepfake-Videos, die in den Kameraeingang der Videokonferenz eingespeist werden, können aufgrund der Verschönerung nicht erkannt werden
- Deepfake Erkennung kann auf entsprechende Filter angepasst werden

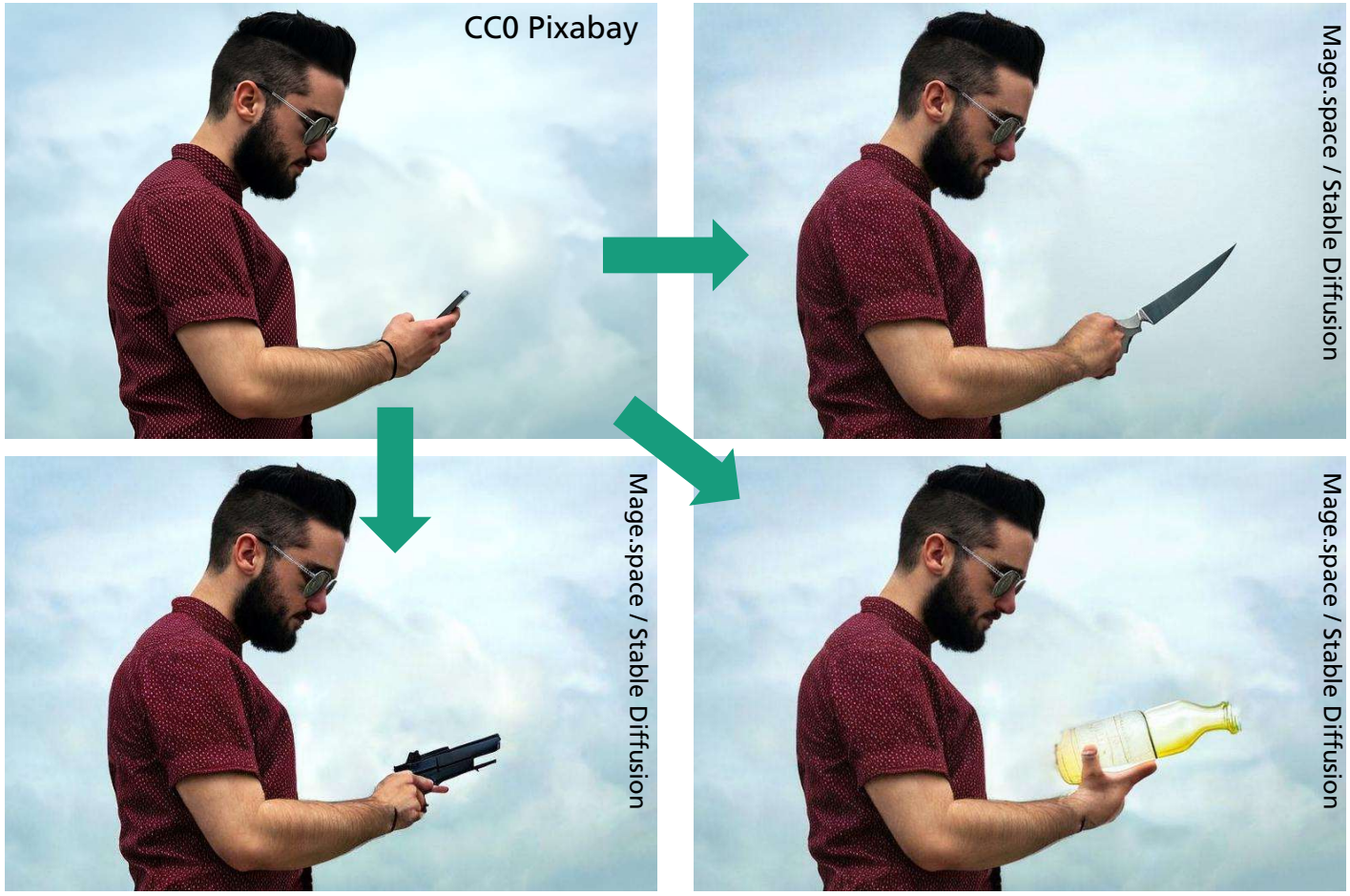
I tried out Zoom's video conferencing both with and without the filter. While the filter doesn't drastically clean up my appearance, the difference is apparent: My skin is made smoother, and the lines and blemishes on my face are softened. Some flyaway hairs are also hidden a touch. You can also notice the filter's effect on the brick wall behind me.

Zoom; Paige Leskin/Business Insider

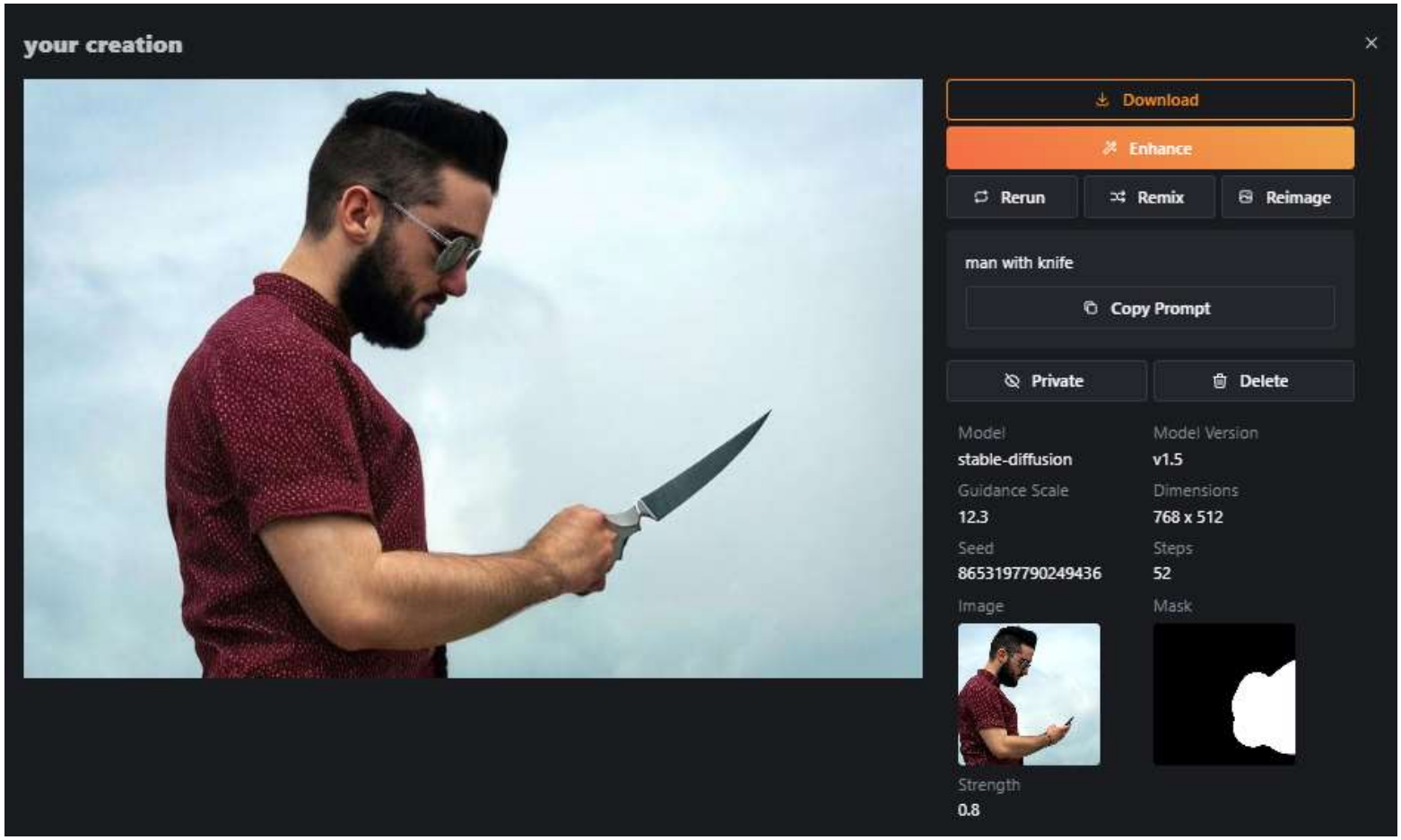
Aktuelle Entwicklung: Text-to-image

- Nächste Generation von ML-basierten Methoden zur Manipulation und Erzeugung von Bilddaten
  - Text2Image: Bilder werden basierend auf einer kurzen Beschreibung generiert
  - Inpainting: Existierende Bilder werden mittels einer kurzen Beschreibung verändert
  - Bekannte Ausprägungen: Dall-E, Stable Diffusion, midjourney, imagen
- Teilweise öffentlich als Web Service verfügbar, teilweise auch open source
- Training der Systeme basiert auf aus dem Internet gecrawlten Inhalte
  - LAION-5B: mehr als 5 Milliarden Paare aus Bild und Text
- Verhinderung von Missbrauch auf verschiedene Arten eingebaut
  - Eingeschränkte Trainingsdaten
  - Filter bei den Textbeschreibungen
  - Kontrolle der Ausgabebilder
  - Lässt sich insbesondere bei Open Source deaktivieren

Aktuelle Entwicklung: Text-to-image



Text-to-image



Text-to-image

Explosion, fire…

Text-2-Image

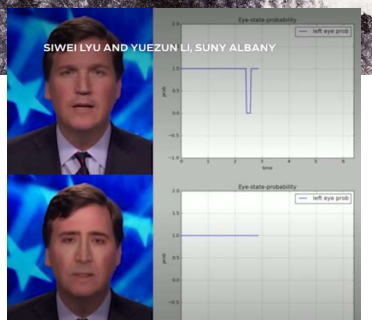
Putin with Sauli Niinisto
(Sputnik/Mikhail Klimentyev/Kremlin
via REUTERS) / infobea

Erkennung über Bildforensische Methoden erscheint möglich

Erkennen von Deepfakes

- Automatisiert durch forensische Methoden
  - Statistische Abweichungen von Bildregionen
  - Abweichungen Audio und Lippenbewegung
- Automatisiert durch maschinelles Lernen
  - Erlernen typischer Eigenschaften von Deepfakes
- Manuell durch Verhaltensmuster
  - Puls, Sprechmuster, Bewegungen
- Semimanuell
  - Erkennen von Quellmaterial durch inverse Bildersuche

https://www.youtube.com/watch?v=ttGUiwfTYvg



Untypische **Gesichtsausdrücke**

Blinzeln

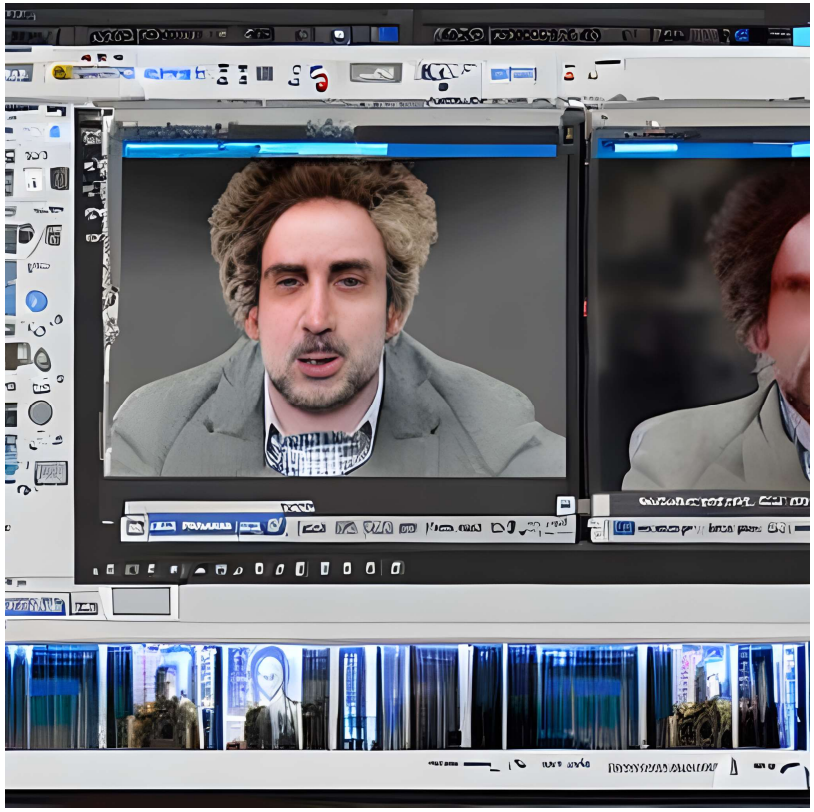
Herzschlag

https://www.youtube.com/watch?v=Ex83dhTn0IU&t=301s

VIELEN DANK.

- Rückfragen sind willkommen.

Kontakt: martin.steinebach@sit.fraunhofer.de



computer running deepfake, stable diffusion