Mediale Identitäten
Angriffsmöglichkeiten und Schadenspotentiale

Dominique Dresen, **Matthias Neu**, Markus Ullmann

Bundesamt
für Sicherheit in der
Informationstechnik

Deutschland
Digital•Sicher•BSI•

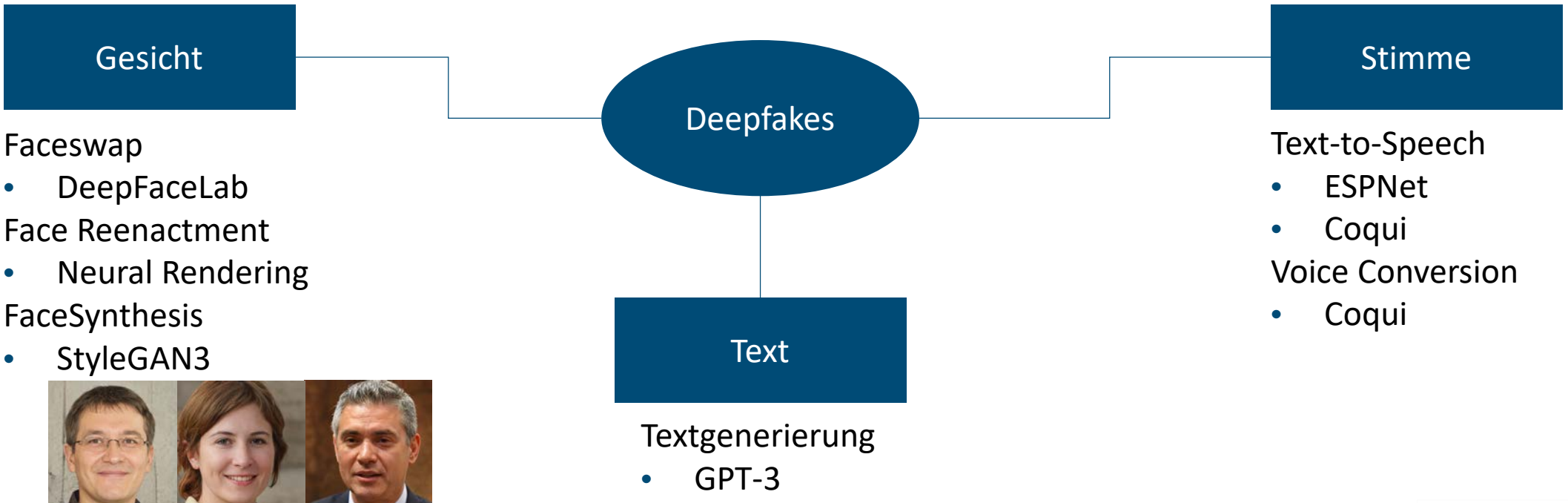
Mediale Identitäten

- Kommunikation mit Mikrofon/Kamera Alltag
- Mediale Identität
  - Repräsentation eines Individuums in einem digitalen Medium
  - Anhand biometrischer Merkmale (Stimme / Gesicht) identifizierbar
- Komplexe und hochdimensionale Daten: bisher nicht leicht zu manipulieren – intuitiv hohes Vertrauen in Authentizität

- Durch Fortschritte im KI-Bereich
  - neue Werkzeuge
  - zunehmend einfacher

Deepfakes – Modalitäten und Methoden

- Deepfake = Deep Learning + Fake
- Umfasst Methoden/Werkzeuge, die mittels KI digitale Identitäten manipulieren können

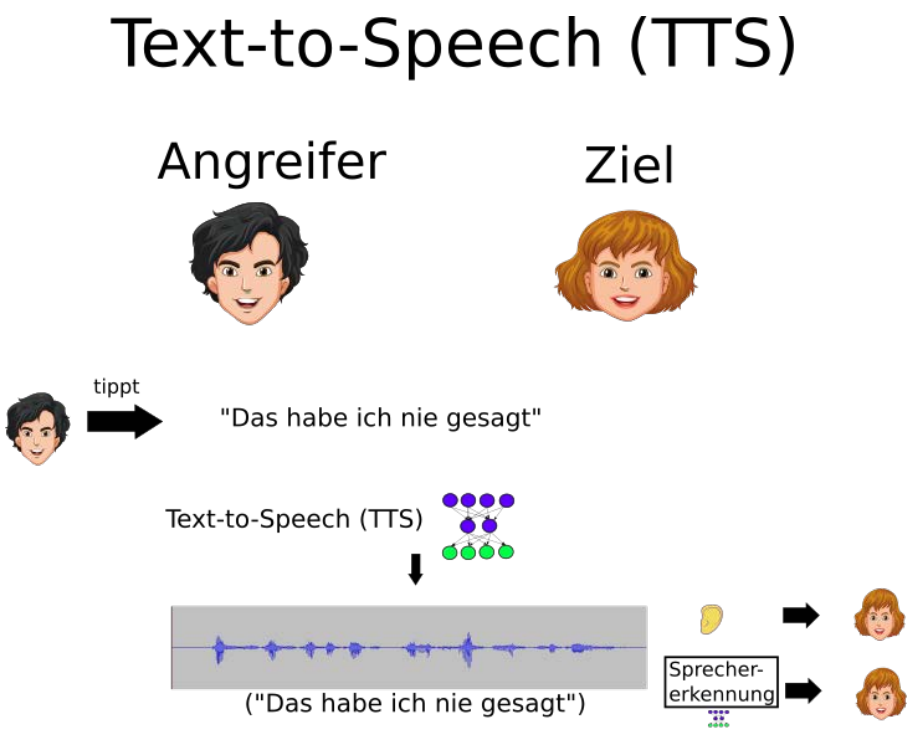


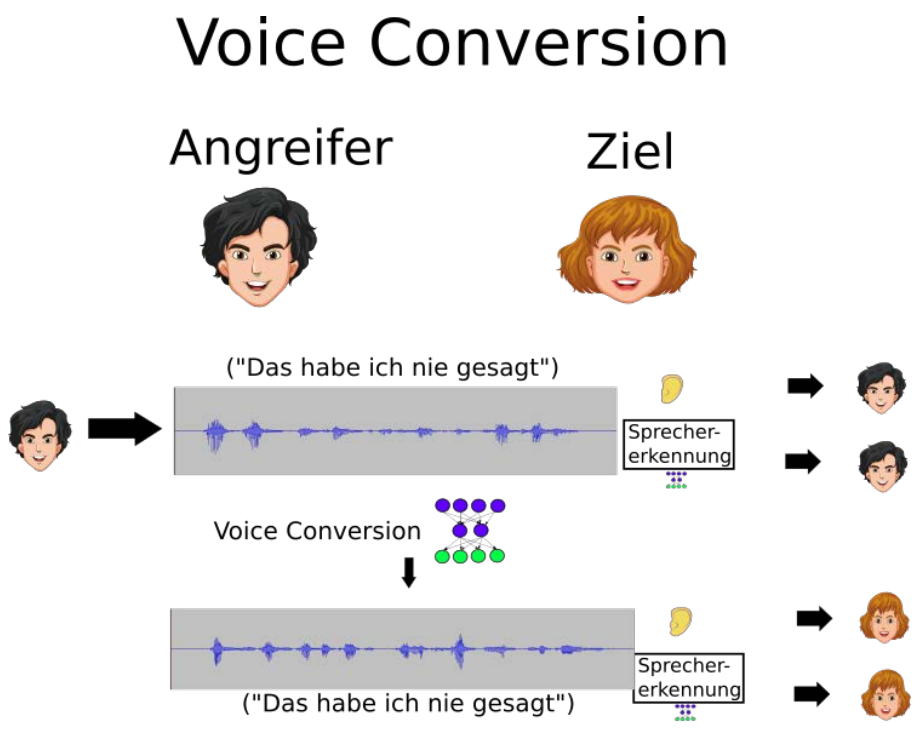
Angriffe auf die Videokommunikation

- **Face-Swap**: Das Gesicht des Angreifers wird ausgetauscht
- **Face-Reenactment:** Die Mimik in einem existierenden Video einer Zielperson wird durch den Angreifer manipuliert
- **(Synthetische Gesichter:** Erzeugung von Gesichtsbildern von Pseudoidentitäten)

Beispiel eines Face-Swaps mit unähnlichen Gesichtern

Angriffe auf die Audiokommunikation





Designed by brgfx / Macrovector / Freepik

Technisches Vorgehen bei Gesichts / Stimmmanipulation

Entwicklung
Softwareentwicklung (Design der neuronalen Netze, Datenvor-/nachbereitungen, Benutzeroberfläche)
Vortrainierung
Benötigt Expertise, ist Zeit- und Rechenaufwendig, kann dann wiederverwertet werden, auch von Laien

Extraktion

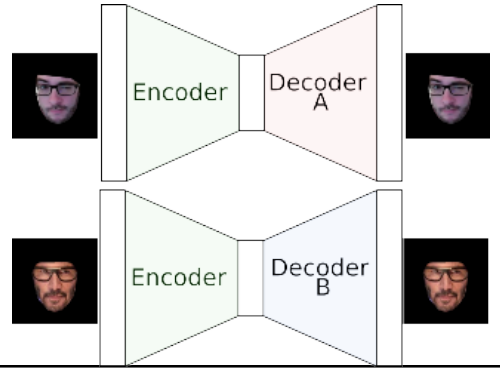
Sammlung und Vorverarbeitung von Daten der Zielperson

≈1.000 – 10.000 Gesichtsbilder
(ca. 5 Minuten Video)
≈min. 5-10 Minuten Audio

Training

Fine-tuning eines vortrainierten Modells mit gesammelten Daten

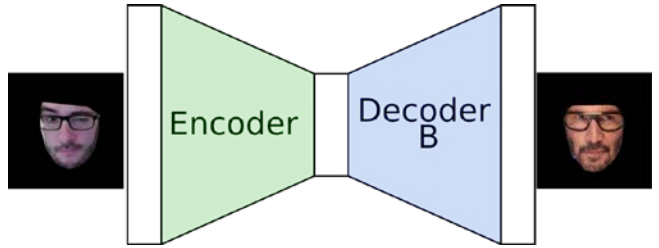
Trainingszeit mehrere Stunden



Inferenz

Einsatz des Modells zur Erzeugung Gesichter oder Audiodaten der Zielperson

Teilweise in Echtzeit möglich



Beispiel (Face-Swap)

Beispiel (Audio)

Originalaufnahme

Fälschung:
Text-to-Speech

Fälschung:
Voice Conversion
mit der Stimme
von Arne Schönbohm

Angriffs- und Bedrohungsszenarien

THE WALL STREET JOURNAL.

Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case

Scams using artificial intelligence are a new challenge for companies

PHOTO: SIMON DAWSON/BLOOMBERG NEWS

By Catherine Stupp
Updated Aug. 30, 2019 12:52 pm ET

SHARE TEXT

Criminals used artificial intelligence-based software to impersonate a chief executive's voice and demand a fraudulent transfer of €220,000 ($243,000) in March in what cybercrime experts described as an unusual case of artificial intelligence being used in hacking.

Forbes

EDITORS' PICK | Oct 14, 2021, 07:01am EDT

Fraudsters Cloned Company Director's Voice In $35 Million Bank Heist, Police Find

TLP:WHITE

Private Industry Notification

FEDERAL BUREAU OF INVESTIGATION, CYBER DIVISION

10 March 2021

PIN Number
210310-001

Please contact the FBI with any questions related to this Private Industry Notification at either your local **Field Office**.

Local Field Offices:
www.fbi.gov/contact-us/field-offices

The following information is being provided by the FBI, with no guarantees or warranties, for potential use at the sole discretion of recipients to protect against cyber threats. This data is provided to help cyber security professionals and system administrators guard against the persistent malicious actions of cyber actors. This PIN was coordinated with DHS-CISA.

This PIN has been released TLP:WHITE: Subject to standard copyright rules, TLP:WHITE information may be distributed without restriction.

Malicious Actors Almost Certainly Will Leverage Synthetic Content for Cyber and Foreign Influence Operations

Summary

Malicious actors almost certainly will leverage synthetic content for cyber and foreign influence operations in the next 12-18 months. Foreign actors are currently using synthetic content in their influence campaigns, and the FBI anticipates it will be increasingly used by foreign and criminal cyber actors for spearphishing and social engineering in an evolution of cyber operational tradecraft.

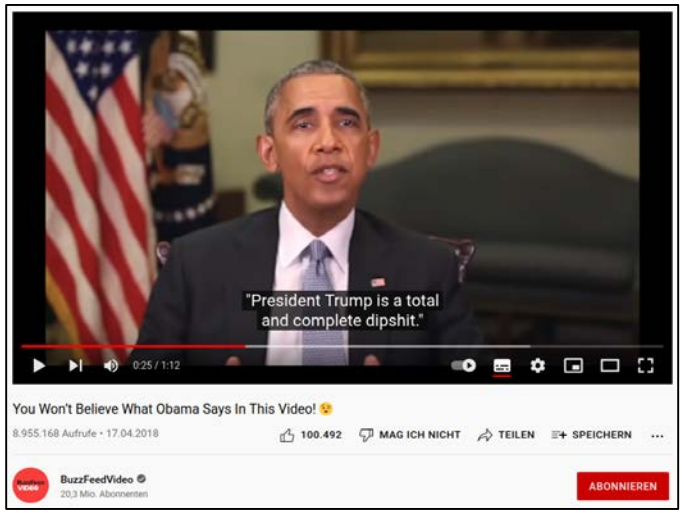
- Vereinzelnd Angriffe in letzten 2 Jahren bekannt geworden
- Deepfake Technologie noch jung, Digitalisierung schreitet voran

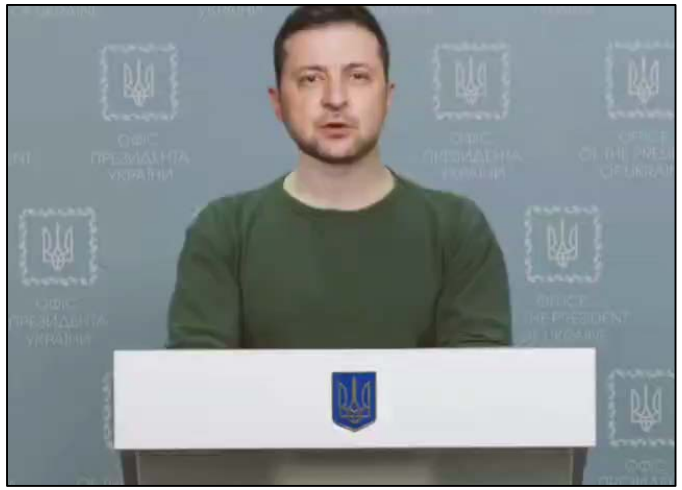
→ Erhöhung der Angriffsfrequenz erwartet

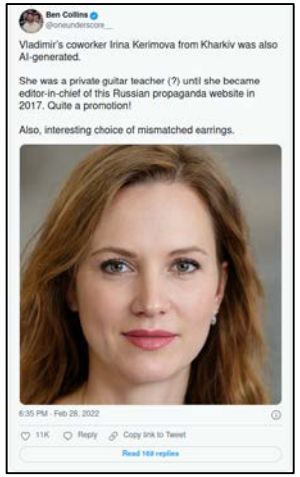
Bundesamt für Sicherheit in der Informationstechnik

Deutschland Digital•Sicher•BSI•

Meinungsmanipulation





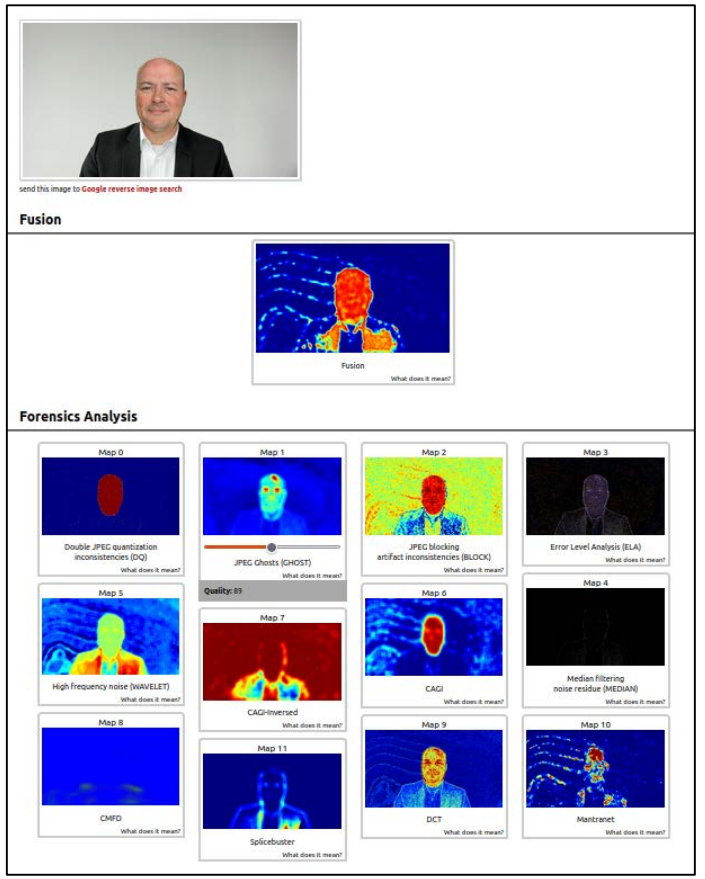


- Manipulation öffentlicher Meinungen durch gefälschte Nachrichten von zentralen Personen der Öffentlichkeit
- Hierzu gehören gefälschte Medieninhalte + Verbreitung über als von der Zielgruppe authentisch wahrgenommene Kanäle
- Nicht nur via Deepfakes, auch „klassische“ Medienmanipulationen

Gegenmaßnahmen

Prävention

- Aufklärung & Awareness
- Multi-Faktor-Authentisierung
- Mehr-Personen-Authorisierung
- Authentizitätsnachweise: Kryptographische Sicherung
- Gesetzliche Regelung
- …



Medienforensik

- Manuell Manipulationen, z.B. auf Pixelebene detektieren
- Soft Biometrics / Verhaltensbasierte Biometrie

Detektion

- KI-Ansätze → Klassifikatoren die auf Deepfake Datenbanken trainiert wurden

Herausforderung der (autom.) Gegenmaßnahmen (1)

Generalisierbarkeit

- Deepfake wurde z.B.
  - mit anderer Methode erstellt, als beim Training der Verteidigungs-KI in Datenbank war
  - auf qualitativ hochwertigen Daten trainiert/evaluiert, aber Angriff wird mit niedriger Auflösung durchgeführt
- Deepfake Detection Challenge (auf kaggle.com von Facebook veranstaltet, 2020):
  - Genauigkeit d. Sieger: Public Dataset 83 %, Black Box Dataset: 65 %
  - Selbe Beobachtung bei ASVSpoof 2021 (Equal Error Rate 0.1 % → 15.6 % → 30% → 50%)

Interpretierbarkeit

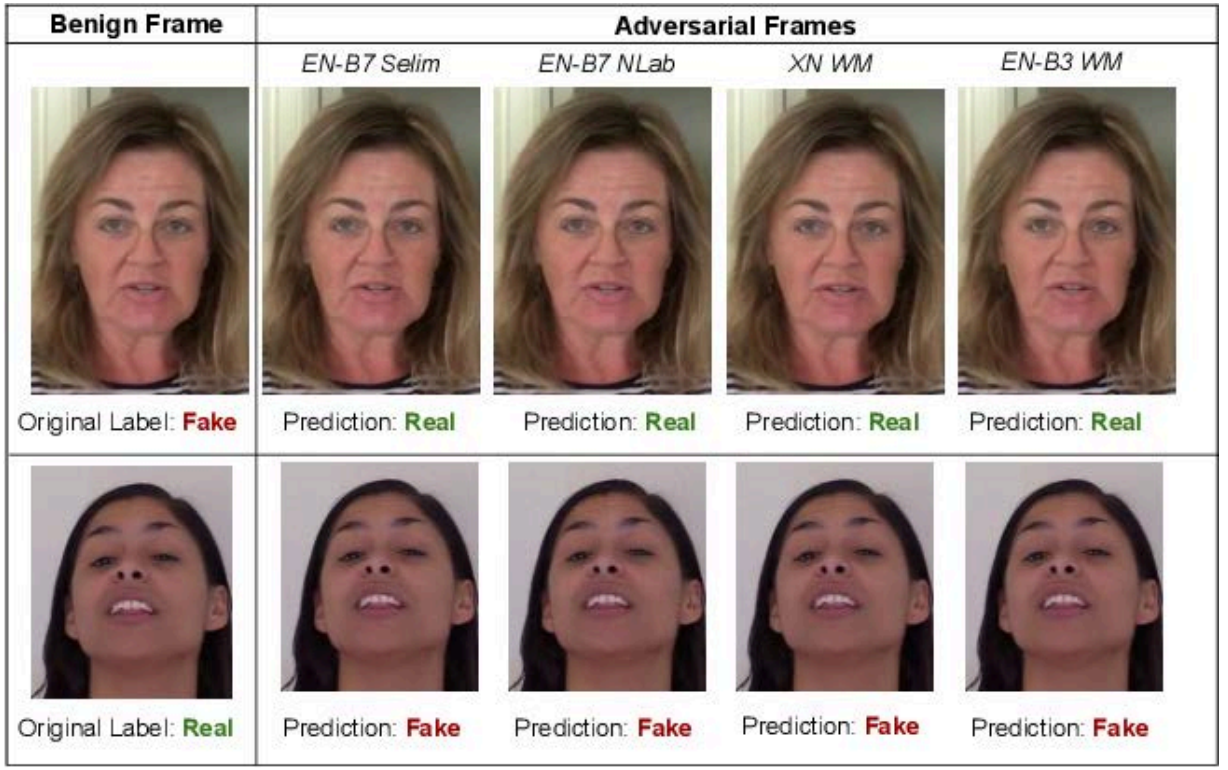
Reaktionsgeschwindigkeit

Herausforderung der (autom.) Gegenmaßnahmen (2)

Adaptive Angreife

Angreifer kennt Faktoren die bei Verteidigung speziell betrachtet werden (z.B. Augen blinken) und intensiviert diese beim Training

Adversariale Angriffe: Angreifer kennt Verteidigungs-KI und nutzt dies um Rauschmuster zu erzeugen, welches die KI umgeht



Beispiel: Adversarialer Angriff auf DF Erkennung. Quelle: [1]

[1] P. Neekhara et al., "Adversarial Threats to DeepFake Detection: A Practical Perspective", 2020

Erzeugung synthetischer Gesichter

- Zur Evaluation biometrischer Verfahren
- Zur Anonymisierung bei der Evaluation von biometrischen System

Ausblick

- Entwicklung der letzten Jahre zeigt die Trends:
  - Steigerung der Qualität
  - Erhöhung der Datenverfügbarkeit zum Erstellen von Modellen
  - Verbesserung der Bedienbarkeit von Werkzeugen zur Erstellung

- Biometrische Systeme zeigen hohe Verwundbarkeit gegen solche Manipulationen

- Alles zusammen lässt häufigeres Auftreten von Gesichts- und Stimmmanipulationen erwarten

- Detektionstechniken müssen gefördert, Präventive Maßnahmen ergriffen werden

Praktische Vorführung

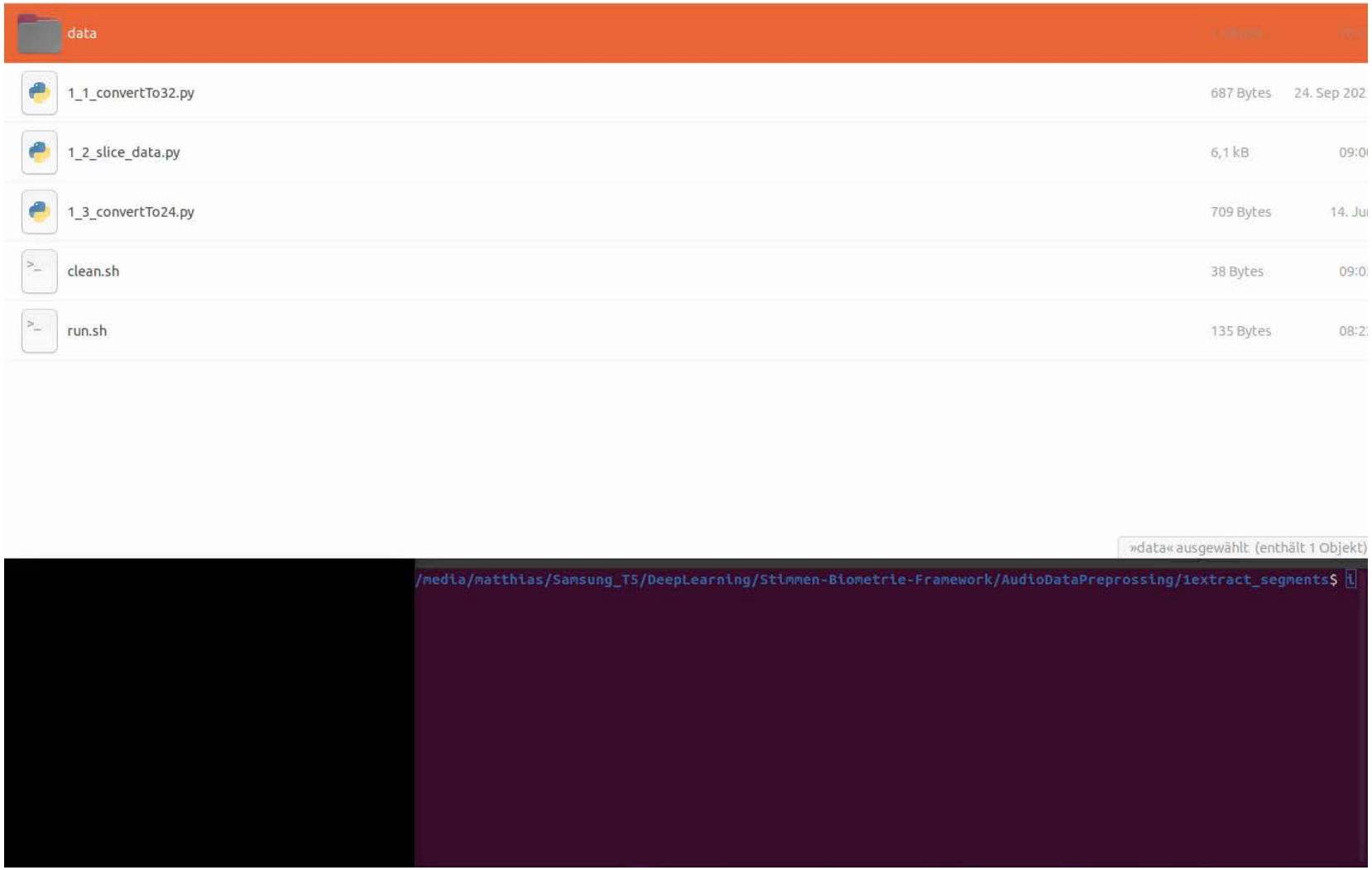
- Interaktive Präsentation von „Deepfake“ Tools und Diskussion

Praktische Vorführung: Klonen einer Stimme

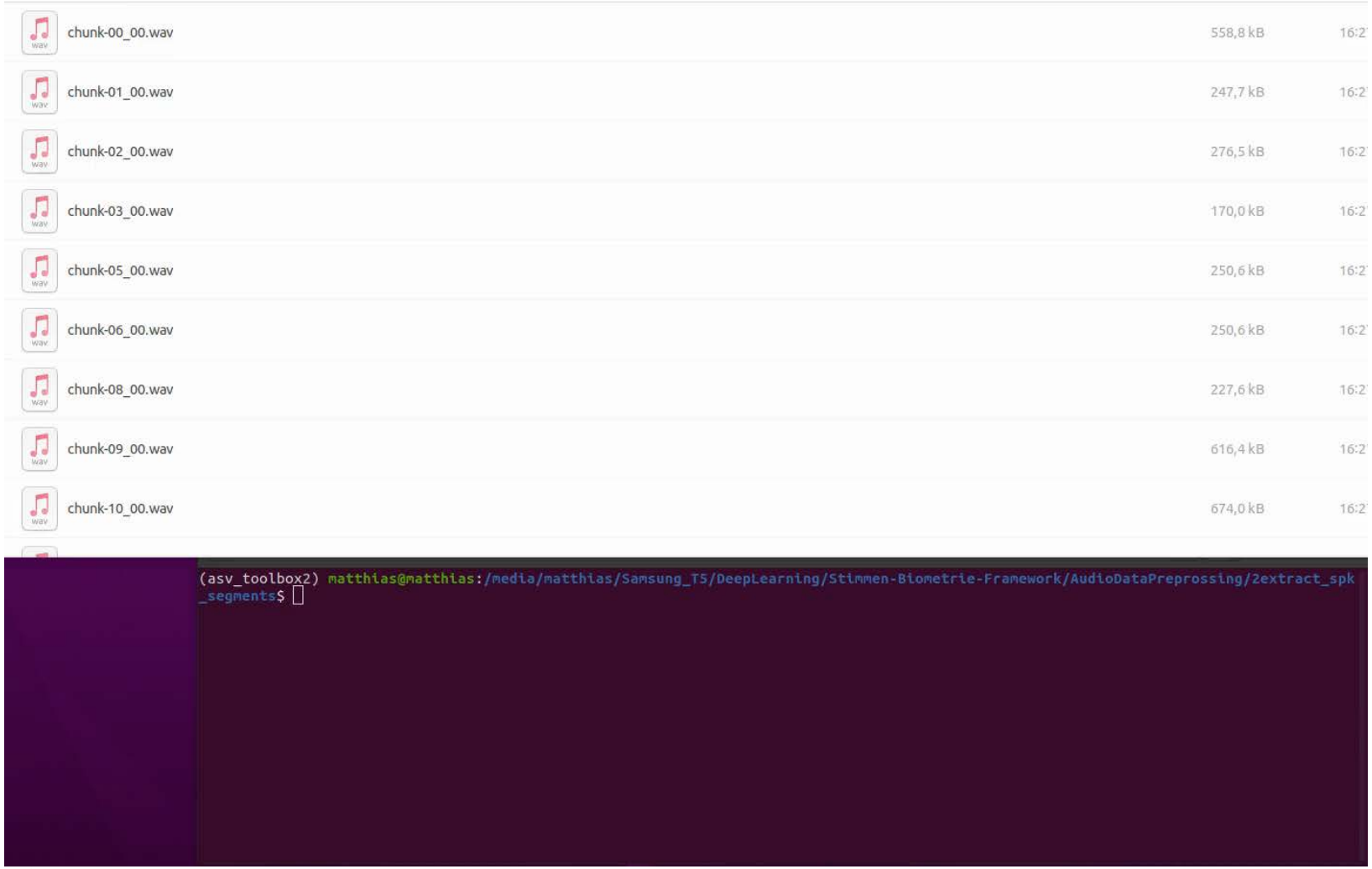
Kriterien:

- Klare Aussprache
- Kein Hall oder andere Störgeräusche
- Nur ein Sprecher
- Möglichst wenige Störungen, wie „ähm“ etc.
- Keine zu langen Pausen innerhalb eines Satzes
- Ähnlicher Stil wie restliche Trainingsdaten (Vorlesen, Rede, Gespräch, etc.)

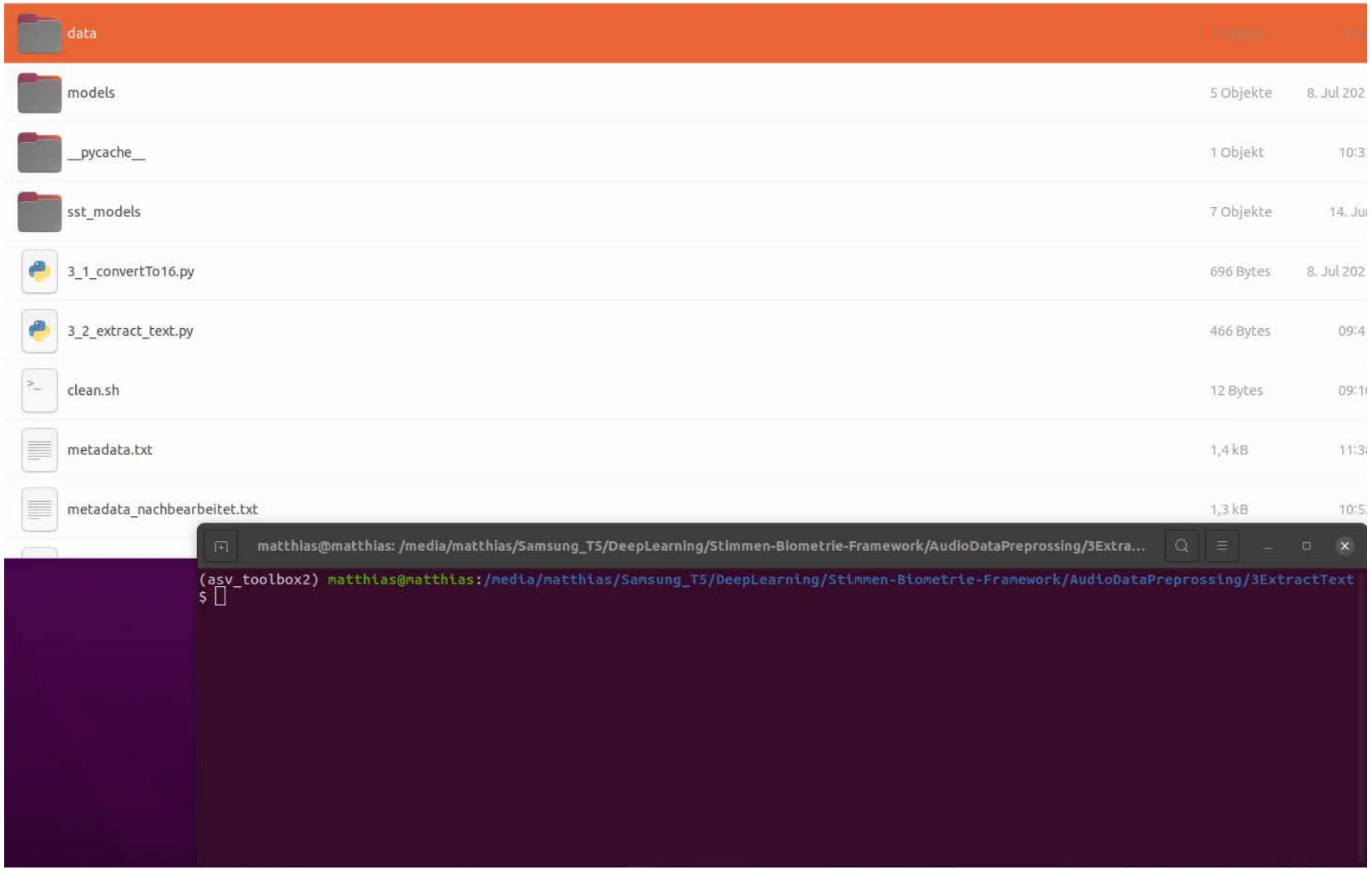
Praktische Vorführung: Aufteilen in einzelne Sätze



Praktische Vorführung: Suche nach dem Zielsprecher



Praktische Vorführung: Transkription



Praktische Vorführung: Inferenz

```
import numpy as np
import IPython
from IPython.display import Audio
import warnings
warnings.filterwarnings('ignore')
import logging
logger = logging.getLogger()
logger.setLevel(logging.CRITICAL)

from asv_toolbox.lib.global_config import global_config
from asv_toolbox.lib.controller.TTS_inference_demonstrator import TTS_inference_Demonstrator

base_path_model = global_config["Model_dir"]
base_path_data = global_config["Data_dir"]
```

Laden des vortrainierten Modells

In [ ]:
```
%%capture
tts_inference_demonstrator = TTS_inference_Demonstrator()
tts_service = "Vitsmelspec_TTS"
model_id_not_trained = "libri_HUI"
model_id = "VCC2018_BSIomni"

#Load the model
tts_inference_not_trained, _ = tts_inference_demonstrator.init_tts_model(tts_service, model_id_not_trained)
```

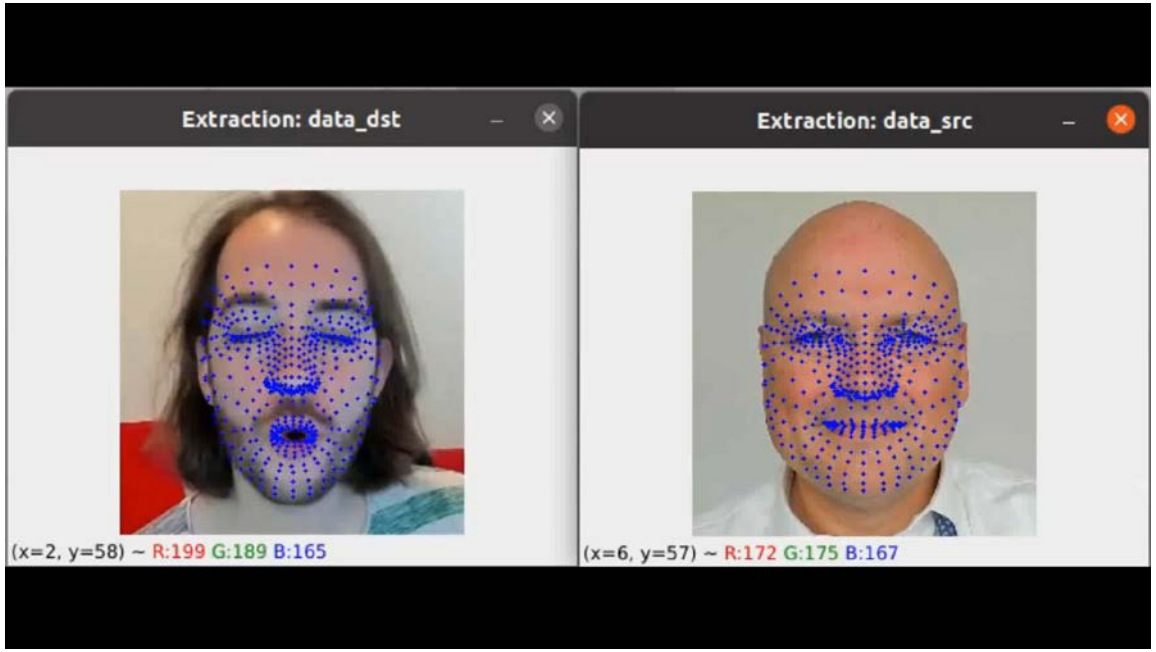
Welcher Text soll generiert werden?

In [ ]:
```
text = "Guten Tag meine Damen und Herren. Ich bin nicht echt. Noch können Sie das wahrnehmen. Zukünftig, mit zunehmen
language = "de-de"
#text = "Dear Ladies and Gentlemen. I am not real and you can still perceive this. But with the increasing maturity o
#language = "en"
```

Audio für Arne Schönbohm

In [ ]:
```
tts_inference_not_trained.set_speakers(target_speaker="BSI_bsi_tts_omnisecure++BSITTS_BSI_AS")
reference_wav_path = tts_inference_not_trained.get_target_wav_path()
y, sr = librosa.load(reference_wav_path, sr=16000)
print("Referenz Audio:")
IPython.display.display(Audio(y, rate=16000))
```

Praktische Vorführung: Gesichtsextraktion



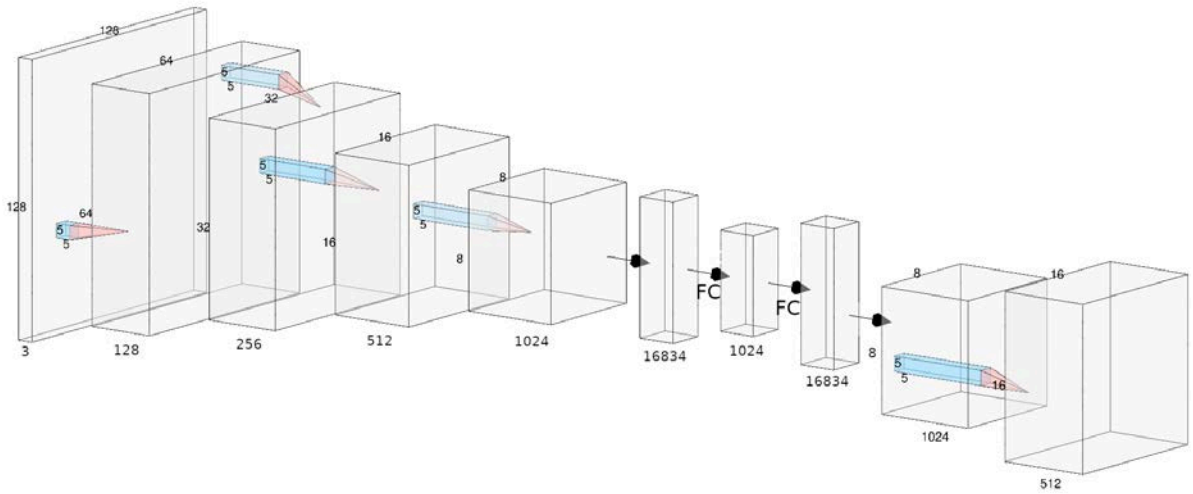
- Extraktion von tausenden ausgerichteten Gesichtsbildern aus Videos von „Angreifer“ und „Zielidentität“

Praktische Vorführung: FaceSwap, einfacher Autoencoder

```python
class Original_Encoder(nn.Module):
    """
    Encoder for the "Original" model:
     - Four successive convolutional layers (each halfing H, W)
     - -> H=W=RES / 2^4, C=1024
     - followed by fc layer with a 1024-dim bottleneck vector
     - First decoding level is shared within the decoder (one upscale layer)
     -> H,W,C = H/8, W/8, 512
    """
    def __init__(self, face_resolution: int):
        super().__init__()
        self.latent_img_res = face_resolution // 2**4
        self.conv1 = nn.Conv2d(3, 128, kernel_size=5, stride=2, padding=2)
        self.conv2 = nn.Conv2d(128, 256, kernel_size=5, stride=2, padding=2)
        self.conv3 = nn.Conv2d(256, 512, kernel_size=5, stride=2, padding=2)
        self.conv4 = nn.Conv2d(512, 1024, kernel_size=5, stride=2, padding=2)
        self.flatten = nn.Flatten()
        self.fc1 = nn.Linear(1024 * self.latent_img_res**2, 1024)
        self.fc2 = nn.Linear(1024, 1024 * self.latent_img_res**2)
        self.upscale = Upscale(1024, 512, scale_factor=2)

    def forward(self, x):
        x = F.leaky_relu(self.conv1(x), 0.1)
        x = F.leaky_relu(self.conv2(x), 0.1)
        x = F.leaky_relu(self.conv3(x), 0.1)
        x = F.leaky_relu(self.conv4(x), 0.1)
        x = self.fc1(self.flatten(x))
        x = self.fc2(x)
        x = x.view((-1, 1024, self.latent_img_res, self.latent_img_res))
        x = self.upscale(x)
        return x
```

- Gesichtsbilder (128x128x3 RGB Bild) werden genutzt um einen für Identitäten geteilten Encoder zu trainieren

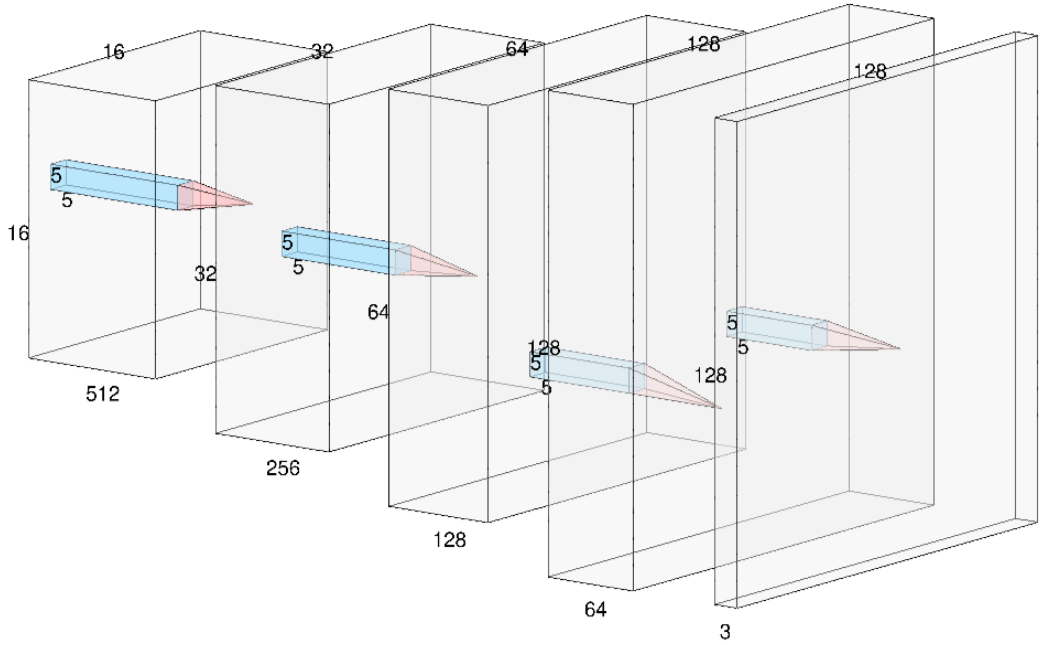


Praktische Vorführung: FaceSwap, einfacher Autoencoder

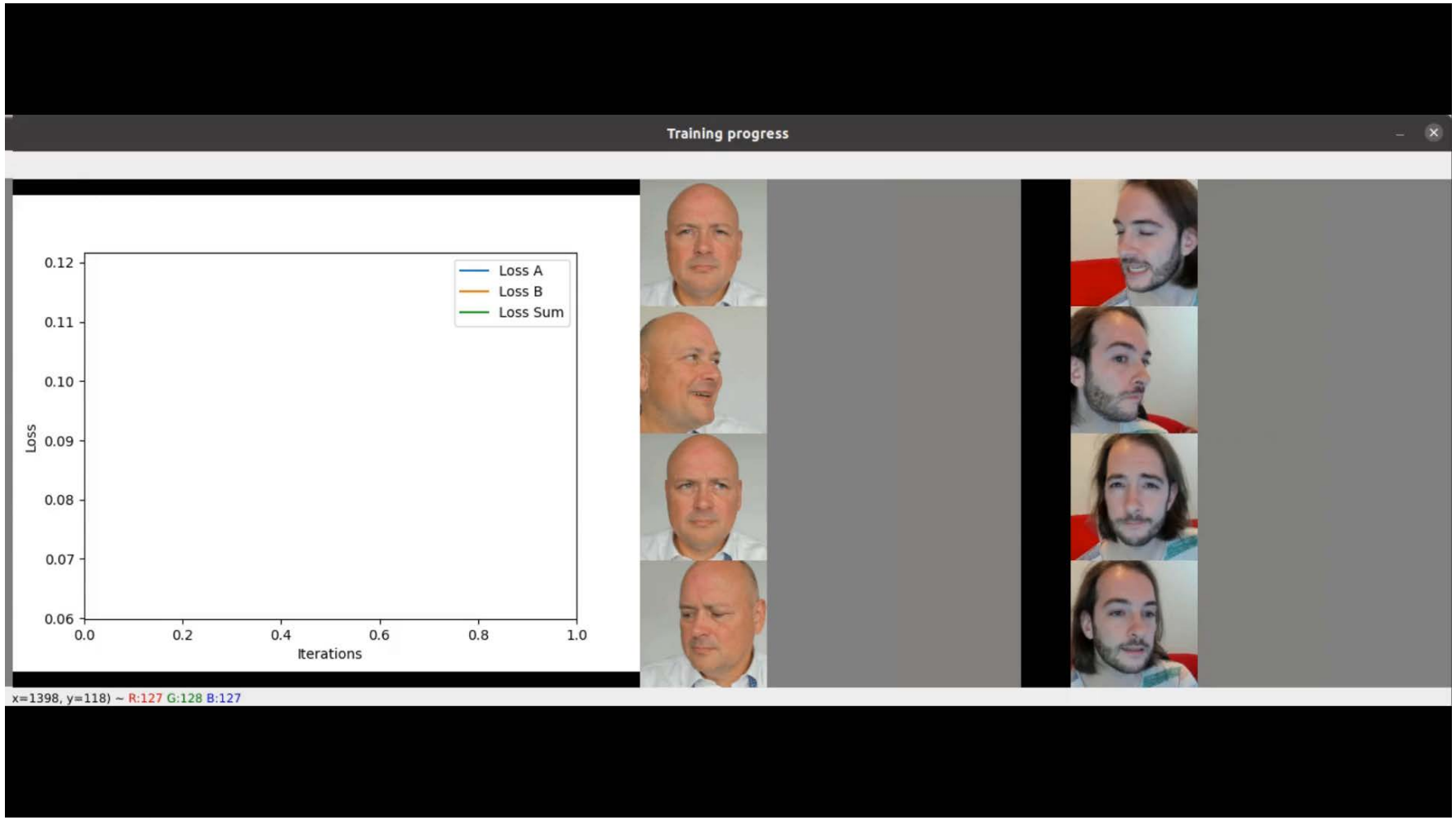
```python
class Original_Decoder(nn.Module):
    """
    Decoder of the "Original" model
    - Expects as input a (512xLxL) tensor
    - three upscale layers (conv + pixel shuffle, each doubling H, W) -> 64x8Lx8L
    - final conv layer to turn tensor into 3x8Lx8L RGB + sigmoid to have range 0..1
    """
    def __init__(self):
        """
        Expects an input channel dimension of 512xWxW, which is transformed to
        3x8*Wx8*W images"""
        super().__init__()
        self.upscale1 = Upscale(512, 256, scale_factor=2)
        self.upscale2 = Upscale(256, 128, scale_factor=2)
        self.upscale3 = Upscale(128, 64, scale_factor=2)
        self.conv = nn.Conv2d(64, 3, kernel_size=5, stride=1, padding=2)

    def forward(self, x):
        x = self.upscale1(x)
        x = self.upscale2(x)
        x = self.upscale3(x)
        x = torch.sigmoid(self.conv(x))
        return x
```

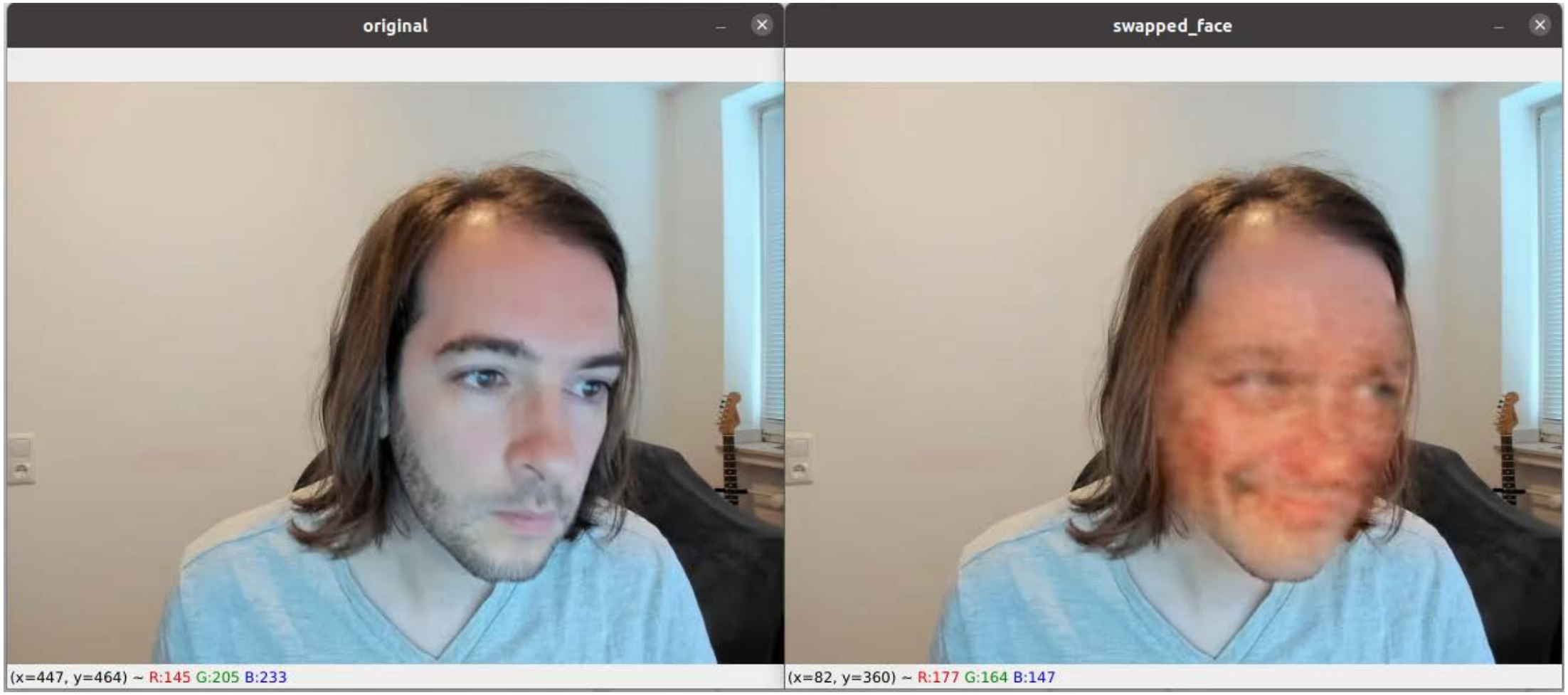
- Training von identitätsspezifischen Decodern um 128x128x3 RGB Bild zu reproduzieren



Praktische Vorführung: FaceSwap, einfacher Autoencoder



Praktische Vorführung: FaceSwap, einfacher Autoencoder



Praktische Vorführung: FaceSwap, Verbesserungen

- Qualität wird signifikant gesteigert durch
  - Datenaugmentation
  - Optimierte Netzwerke + Trainingsroutine
  - Höhere Gewichtung von Augen/Mund Region
  - Zusätzlicher GAN Diskriminator
  - Gesichtsfarbenanpassung
  - …

Praktische Vorführung: FaceSwap, Verbesserungen

Vielen Dank für Ihre Aufmerksamkeit!

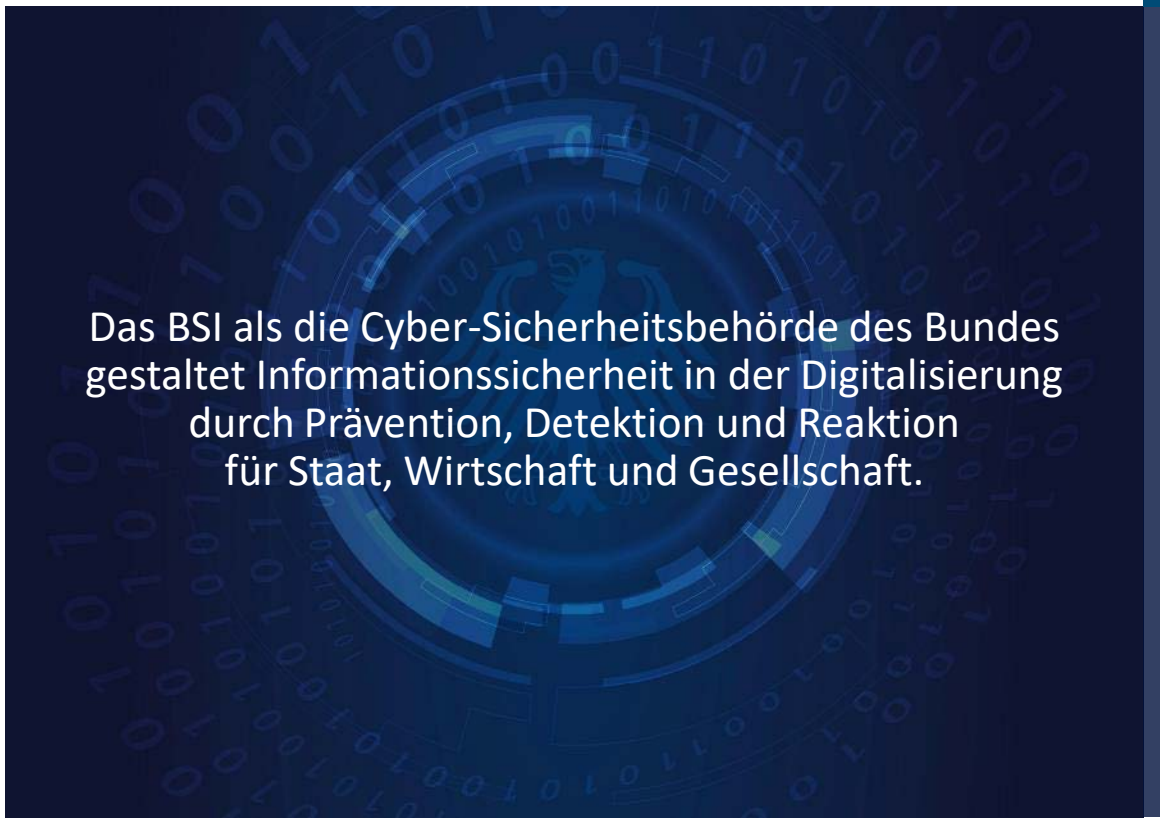
Deutschland
Digital•Sicher•BSI•

Kontakt

Dominique Dresen / Matthias Neu
Referenten

matthias.neu@bsi.bund.de
Tel. +49 (0) 228 9582 - 6834 / 6604

Bundesamt für Sicherheit in der Informationstechnik (BSI)
Godesberger Allee 185-189
53175 Bonn
www.bsi.bund.de



Bundesamt
für Sicherheit in der
Informationstechnik