HHN Face Wizard

Interaktives Deep Fake User Interface

Pascal Graf und Nicolaj C. Stache
Heilbronn University of Applied Sciences

Projektübersicht

Möglichkeiten:

- Der Demonstrator zeigt die Fähigkeit modernster KI-Algorithmen, das Gesicht eines Menschen in einer Filmszene (Target) durch ein beliebiges anderes Gesicht (Source) austauschen, wobei der Gesichtsausdruck der Target-Person erhalten bleibt.
- Zudem ist es mit dem *Face Wizard* möglich, die Mimik einer Person von einem Video auf eine beliebige andere Person zu übertragen.

Ansatz / Ziele:

- Für die *Face Swap* Funktion erstellt ein neuronales Netzwerk Bilder mit der Identität der Source-Person und der Mimik aus dem Target-Image.
- Für das *Face Reenactment* wird die Mimik unabhängig von der Identität aus einem Video extrahiert und durch ein neuronales Netzwerk auf eine andere Person übertragen.

Zukünftige Anwendungsfelder:

- Die verwendete Technologie bietet Möglichkeiten im Bereich automatisierter Übersetzung von (Lehr-)Videos sowie der Synchronisation von Filmen.
- Zudem soll der Demonstrator über Fähigkeiten und Risiken modernster KI-Algorithmen informieren.

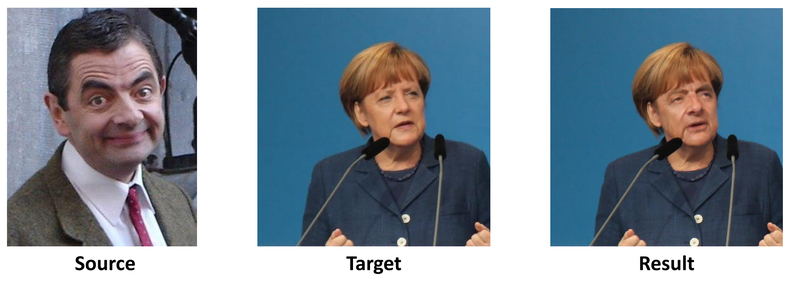

Abbildung 1: Beispiel für Source-, Target- und Result-Image beim Face Swap mit dem Face Wizard [3][4].

Face Swap

Funktionsweise:

- Basiert auf „SimSwap: An Efficient Framework For High Fidelity Face Swapping“ [1].
- Ein *Generator* extrahiert Informationen über Identität und Mimik aus dem Target-Image und ersetzt die Identität durch jene aus einem Source-Image.
- Zum Training werden zusätzlich ein *Diskriminator*, der beurteilt, wie realistisch das Bild ist sowie ein Identity Extractor, der die Identitäten der Personen extrahiert, genutzt.

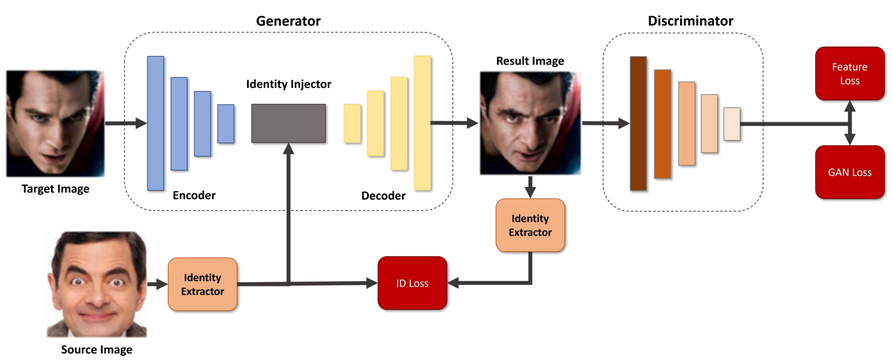

Abbildung 2: Schematische Darstellung von Programmablauf und verwendeten Architekturen in „SimSwap: An Efficient Framework For High Fidelity Face Swapping“. Ein *Generator* extrahiert zunächst Identitäts- und Mimikinformationen aus einem Target Image. Dann wird die Identität einer Person aus dem Source Image hineingespeist und ein neues Result Image generiert. Der *Diskriminator* wird nur für den Trainingsprozess benötigt.

Trainingsprozess:

- Das vom Generator erzeugte Bild soll drei Kriterien erfüllen:
  1. Es soll ein realistisches menschliches Gesicht zeigen.
  2. Die Identität des Gesichts soll mit dem des Source-Images übereinstimmen.
  3. Die Mimik soll mit der des Target-Images übereinstimmen.
- Diese Kriterien werden über entsprechende Loss-Funktionen während des Trainings überprüft und die Netzwerkgewichte dahingehend optimiert.
- Trainiert wird auf Basis von etwa 500.000 Bildern von über 8.600 verschiedenen Personen (VGGFace2).

Abbildung 3: Source-, Target- und Result-Bilder verschiedener Personenkombinationen für einen Face Swap mithilfe der Simswap-Architektur [1].

Face Reenactment

Funktionsweise:

- Basiert auf „Depth-Aware Generative Adversarial Network for Talking Head Video Generation“ [2]. Ziel ist es, die Mimik eines Menschen auf eine andere Person im Target-Image zu übertragen
- Der Generator besteht aus drei Teilen:
  1. Ein *Face Depth Network* erstellt die Tiefenkarte eines Gesichts.
  2. Diese Tiefenkarte wird gemeinsam mit dem Farbbild genutzt, um wichtige Keypoints mit dem *Keypoint Estimator* zu extrahieren.
  3. Features aus einem *Attention Module* werden genutzt, um ein Bild zu generieren, welches das Gesicht aus dem Target-Image, jedoch die Pose des Driving-Images zeigt.

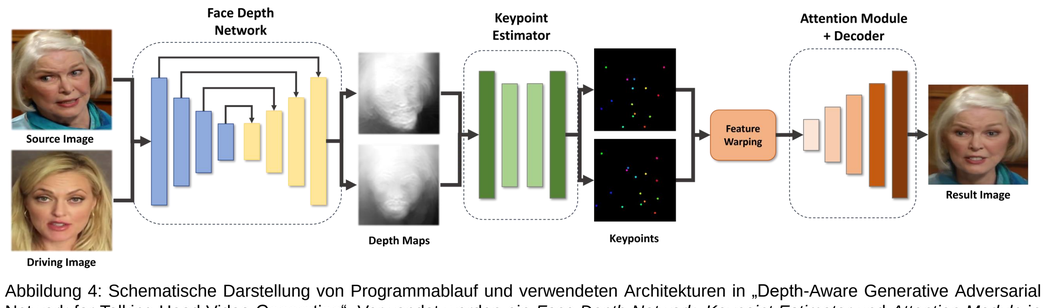

Abbildung 4: Schematische Darstellung von Programmablauf und verwendeten Architekturen in „Depth-Aware Generative Adversarial Network for Talking Head Video Generation“. Verwendet werden ein *Face Depth Network*, *Keypoint Estimator* und *Attention Module* in Kombination mit einem *Decoder*.

Trainingsprozess:

- Das *Depth Network* wird mithilfe zweier aufeinanderfolgender Bilder desselben Videos darauf trainiert, eine Tiefenkarte zu erstellen (*Self-Supervised Training*).
- Das *Attention Module* erlernt, welche Bereiche des Bildes wichtig für die Veränderung der Mimik sind.

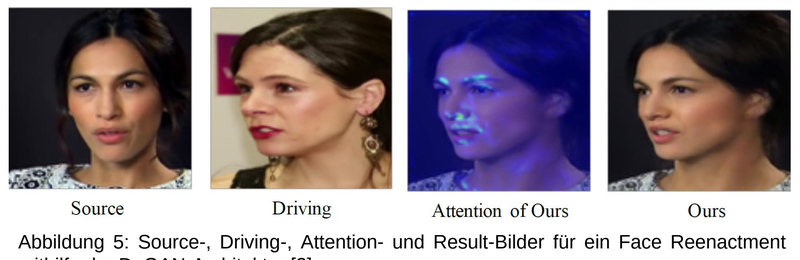

Abbildung 5: Source-, Driving-, Attention- und Result-Bilder für ein Face Reenactment mithilfe der DaGAN-Architektur [2].

Weitere Forschung

- Neben der Generierung und Modifikation von Bildinhalten, können ähnliche Deep Learning Methoden auch zur Nachahmung von Stimmen genutzt werden. Ein Teil unserer Forschungsanstrengungen widmet sich der Integration solcher Techniken in den Demonstrator.
- Wegen der Qualität, die Deep Fakes erreicht haben, untersucht die Hochschule Heilbronn außerdem Methoden, mit denen künstliche erzeugte Inhalte erkannt und Nutzer*innen gewarnt werden können.

Quellen

[1] R. Chen, X. Chen, B. Ni, and Y. Ge, ‘SimSwap: An Efficient Framework For High Fidelity Face Swapping’, CoRR, vol. abs/2106.06340, 2021.

[2] F.-T. Hong, L. Zhang, L. Shen, and D. Xu, ‘Depth-Aware Generative Adversarial Network for Talking Head Video Generation’, arXiv [cs.CV]. 2022.

[3] Antonio Zugaldia from Brussels, Belgium (https://commons.wikimedia.org/wiki/File:Mr._Bean_2011.jpg), „Mr. Bean 2011“, https://creativecommons.org/licenses/by/2.0/legalcode

[4] Indeedous (https://commons.wikimedia.org/wiki/File:Angela_Merkel_Apolda_2014_003.jpg), “Angela Merkel Apolda 2014”, Wikimedia Commons

Pascal.Graf@hs-heilbronn.de, Nicolaj.Stache@hs-heilbronn.de