AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

FAKE NEWS AUF DER SPUR

Ein Forschungsschwerpunkt über Medienmanipulation

Dr. Alexander Schindler

Scientist

Data Science & Artificial Intelligence
Center for Digital Safety & Security
AIT Austrian Institute of Technology GmbH

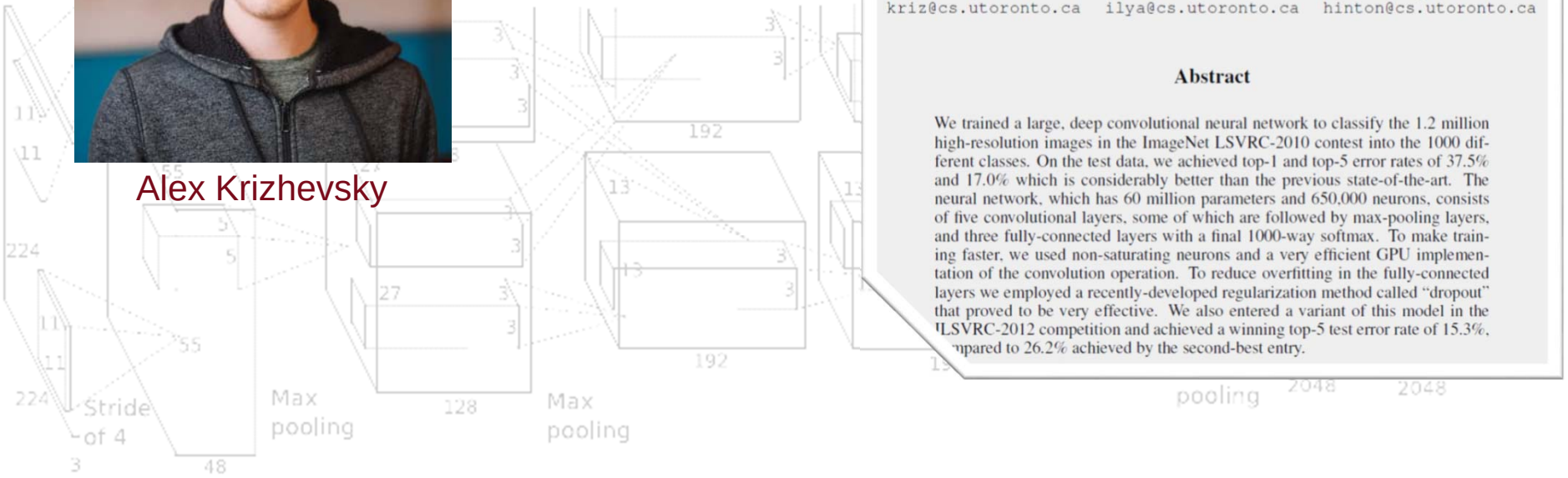
AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

AUFBRUCH!

Motto Austrian Innovation Forum 2020

2012 - AlexNet

Alex Krizhevsky



ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

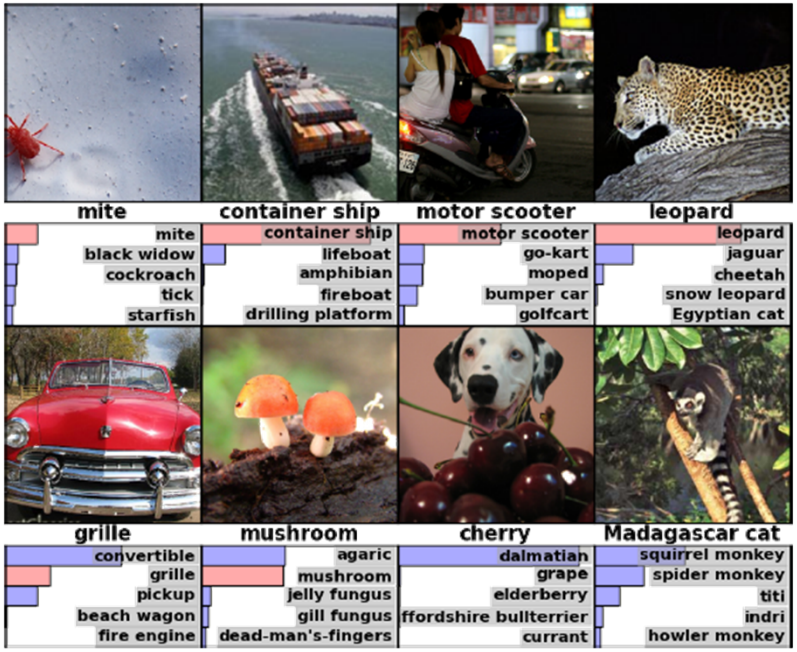
Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

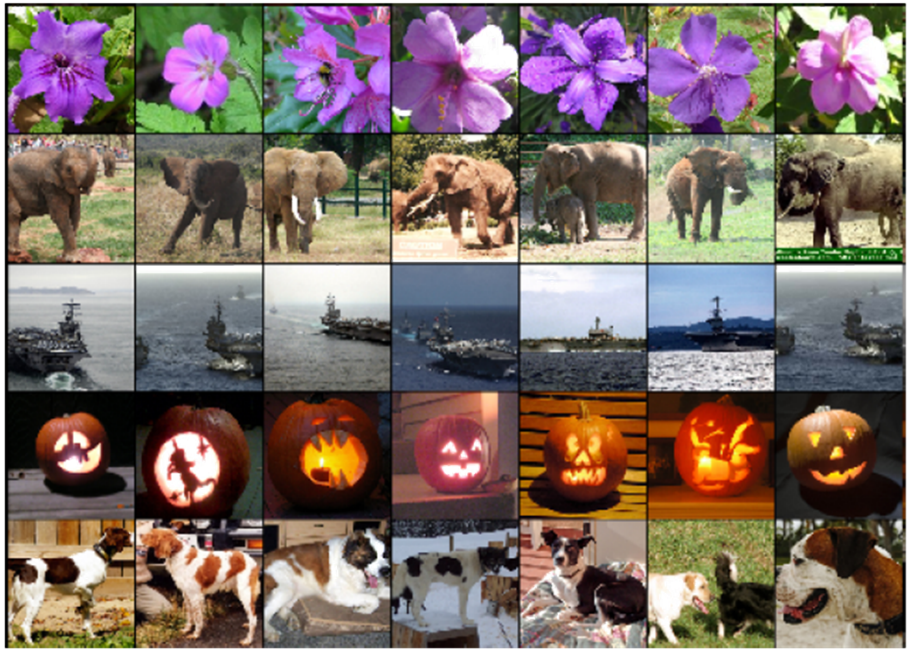
Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

Abstract

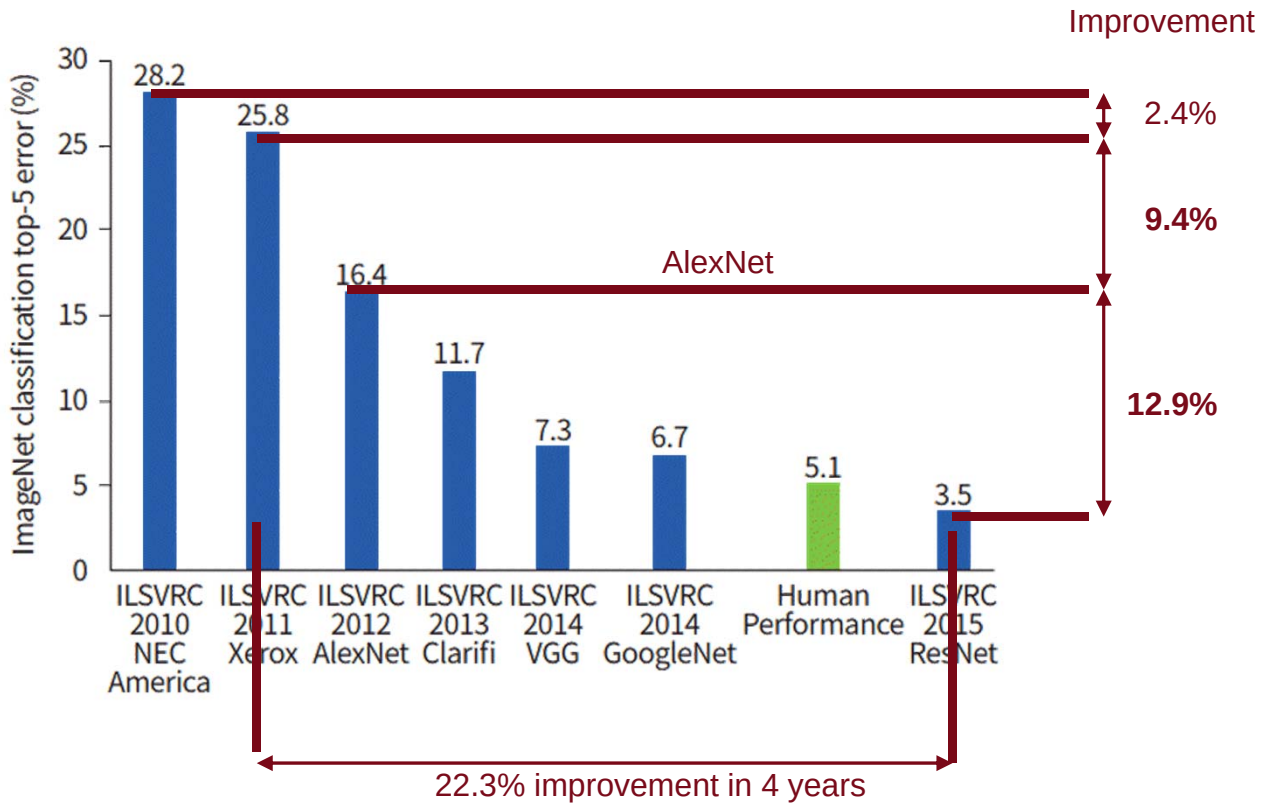
We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called "dropout" that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.

Task – Image Captioning





Dropping error rates since then



Rising charts and trend lines everywhere

- **Science**
  - AI Conferences
- **Politics**
  - AI public funding
  - AI national Strategies
- **Private Sector**
  - Departments renamed to „AI …“
  - Influencers and Enablers
  - AI Events and Summits

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

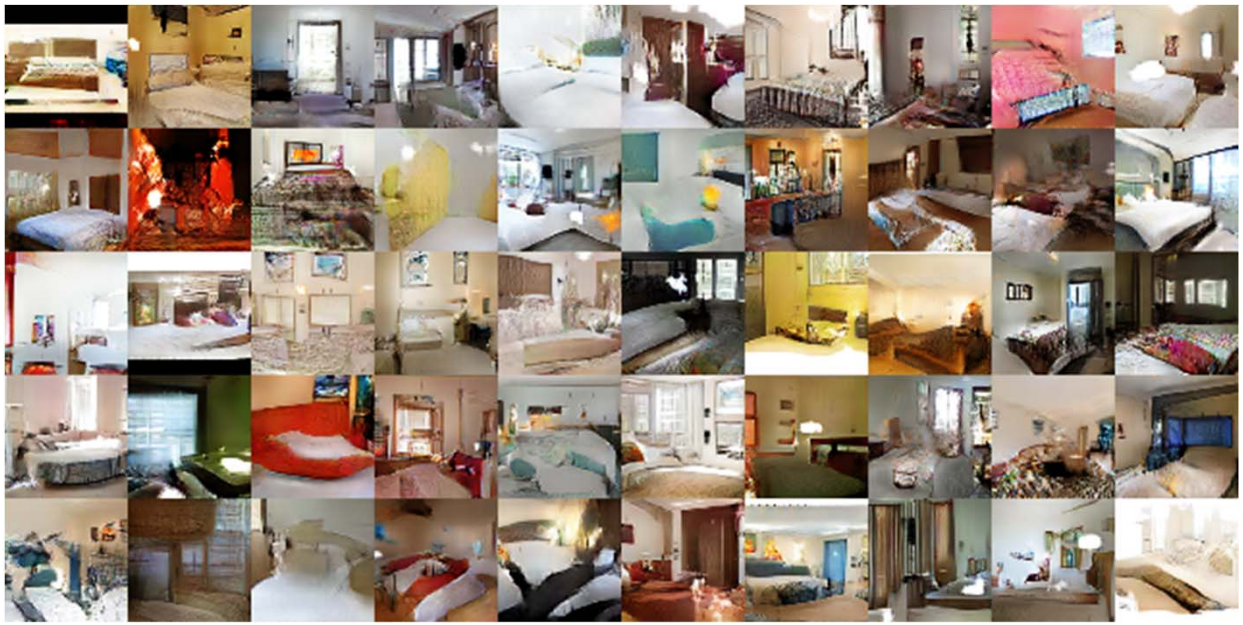
NEUE TECHNOLOGIEN

Neue Möglichkeiten

Image Generation (2015)

Generative Adversarial Networks: One network transforms random noise into images, a second one tries to distinguish generated from real images. Both are trained at the same time.

Nov 2015: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, http://arxiv.org/abs/1511.06434, https://github.com/Newmu/dcgan_code



Photographs of bed rooms that do not actually exist

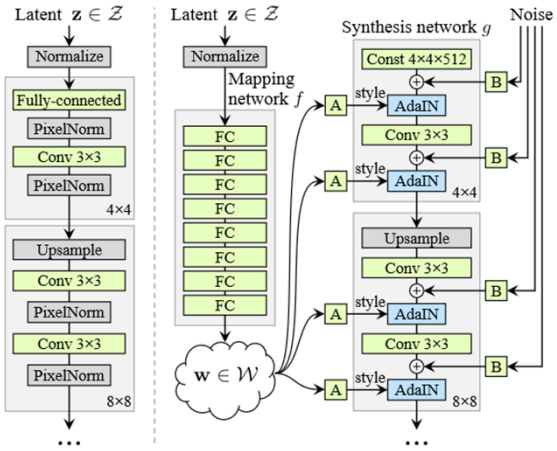
A Style-Based Generator Architecture for Generative Adversarial Networks

Tero Karras
NVIDIA
tkarras@nvidia.com

Samuli Laine
NVIDIA
slaine@nvidia.com

Timo Aila
NVIDIA
taila@nvidia.com

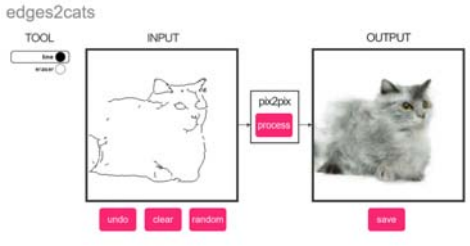
AUSTRIAN INSTITUTE OF TECHNOLOGY

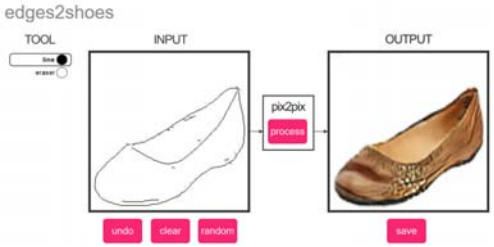


(a) Traditional (b) Style-based generator

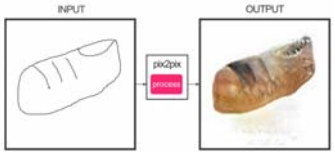
Pix2Pix (2017)

Cherry-picked Examples





My attempts



Tool Web site: https://affinelayer.com/pixsrv/

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros. Image-to-Image Translation with Conditional Adversarial Networks. CVPR, 2017.

NVIDIA GAUGAN

Semantic Image Synthesis with Spatially-Adaptive Normalization

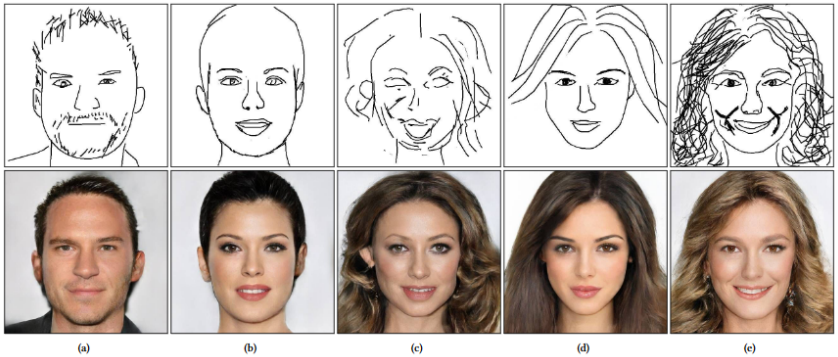
Taesung Park[1,2*] Ming-Yu Liu[2] Ting-Chun Wang[2] Jun-Yan Zhu[2,3]

[1]UC Berkeley [2]NVIDIA [2,3]MIT CSAIL

Park, T., Liu, M. Y., Wang, T. C., & Zhu, J. Y. (2019). Semantic image synthesis with spatially-adaptive normalization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2337-2346).

DeepFaceDrawing: Deep Generation of Face Images from Sketches

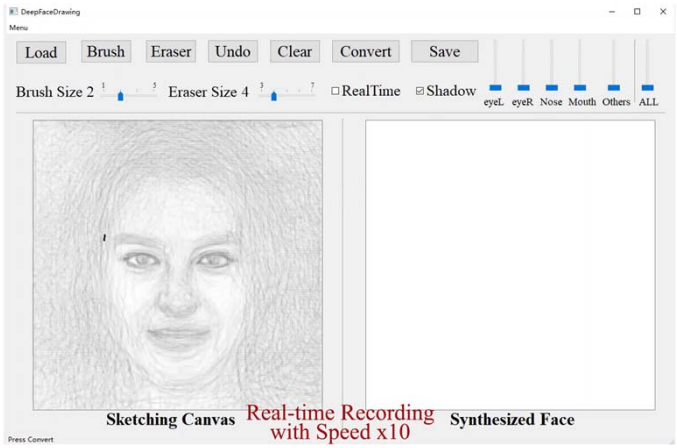
SHU-YU CHEN†, Institute of Computing Technology, CAS and University of Chinese Academy of Sciences
WANCHAO SU†, School of Creative Media, City University of Hong Kong
LIN GAO*, Institute of Computing Technology, CAS and University of Chinese Academy of Sciences
SHIHONG XIA, Institute of Computing Technology, CAS and University of Chinese Academy of Sciences
HONGBO FU, School of Creative Media, City University of Hong Kong

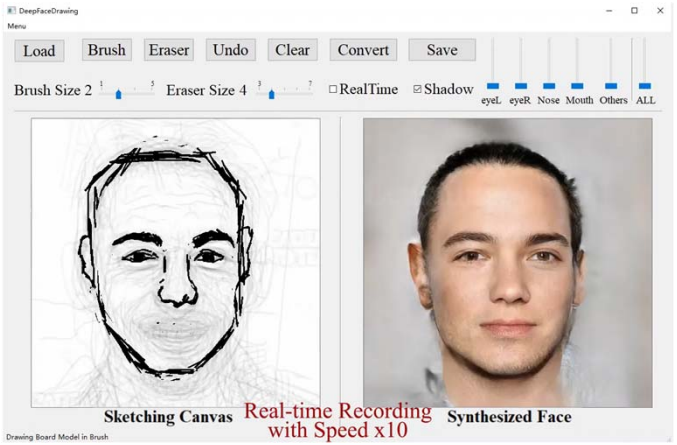


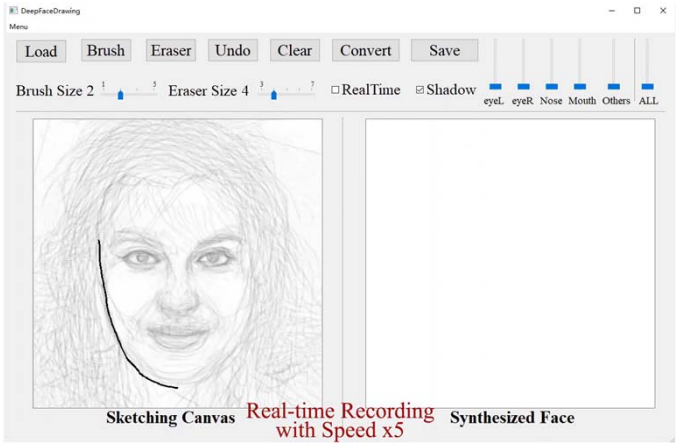
Tool Web site: http://www.geometrylearning.com:3000/index_EN_621.html

Chen, S. Y., Su, W., Gao, L., Xia, S., & Fu, H. (2020). DeepFaceDrawing: deep generation of face images from sketches. ACM Transactions on Graphics (TOG), 39(4), 72-1.

16.10.2020





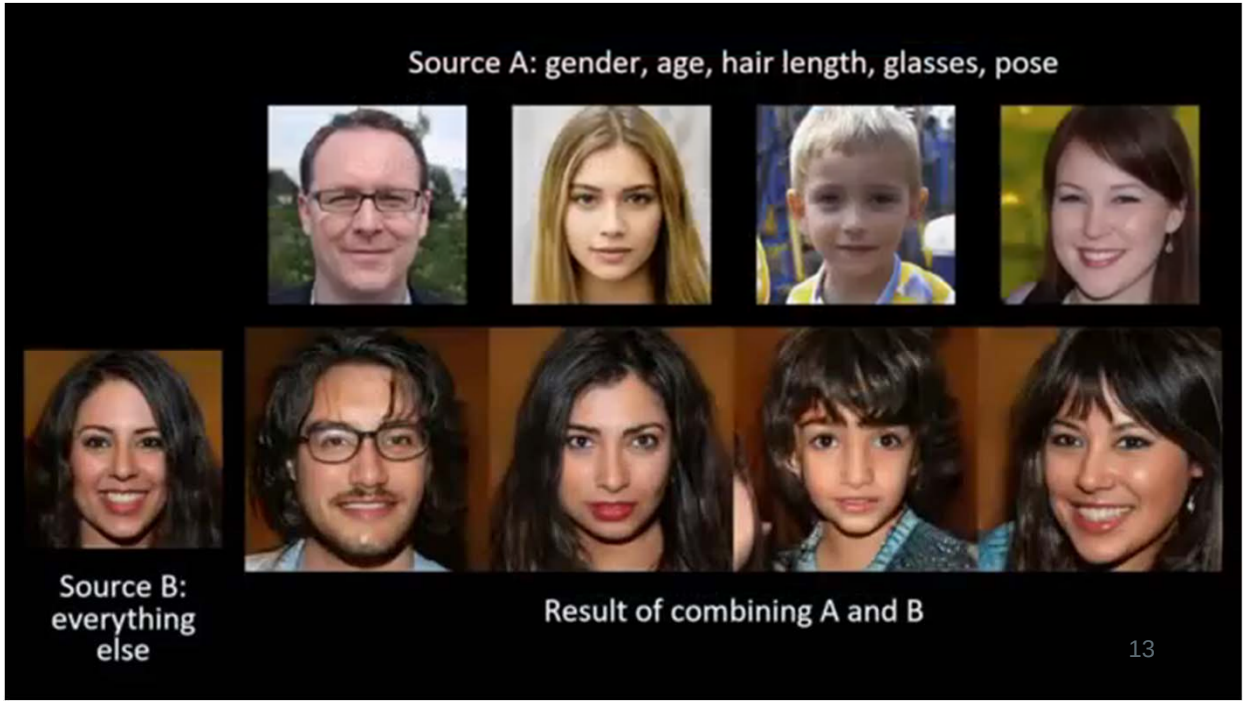


A Style-Based Generator Architecture for Generative Adversarial Networks

Tero Karras
NVIDIA
tkarras@nvidia.com

Samuli Laine
NVIDIA
slaine@nvidia.com

Timo Aila
NVIDIA
taila@nvidia.com



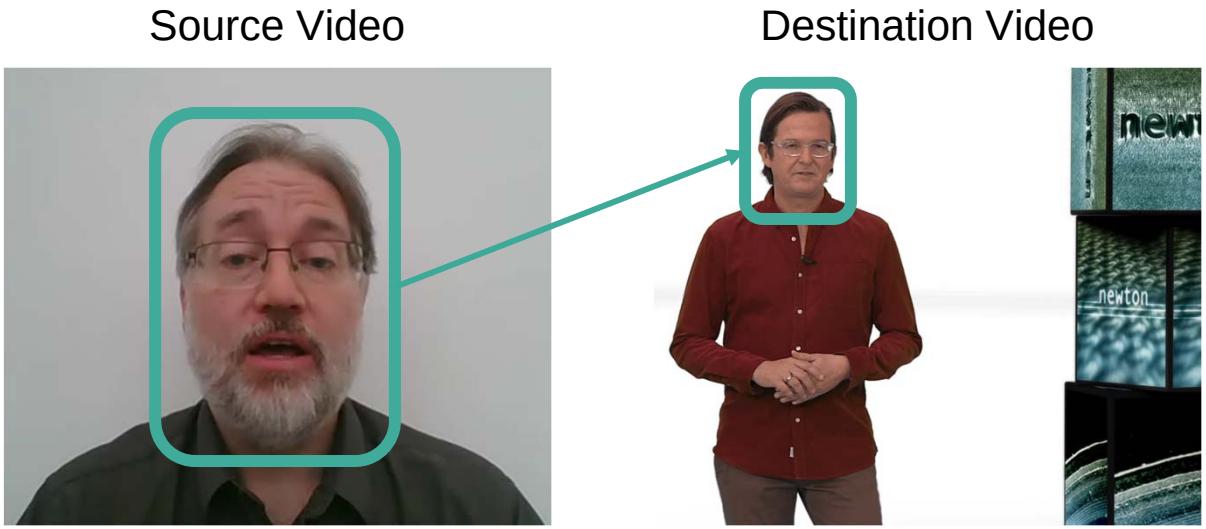
Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4401-4410).

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

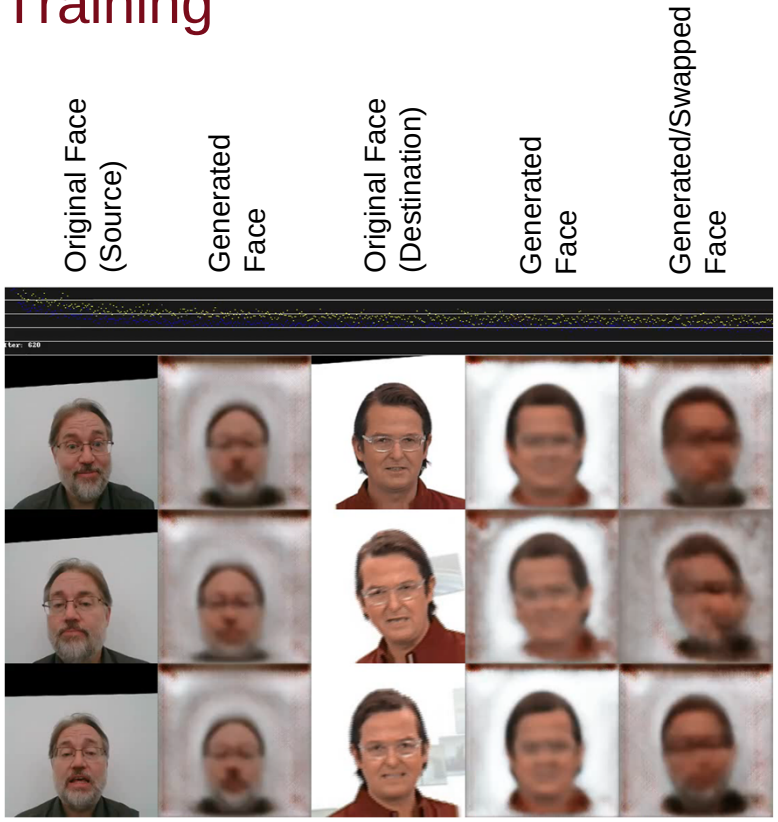
DEEP FAKES

Next Level Content Generation

Face-Swap

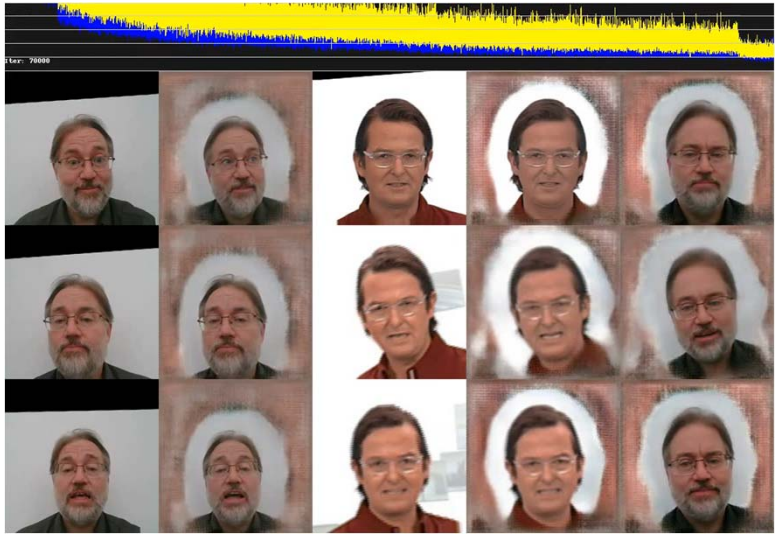


Training

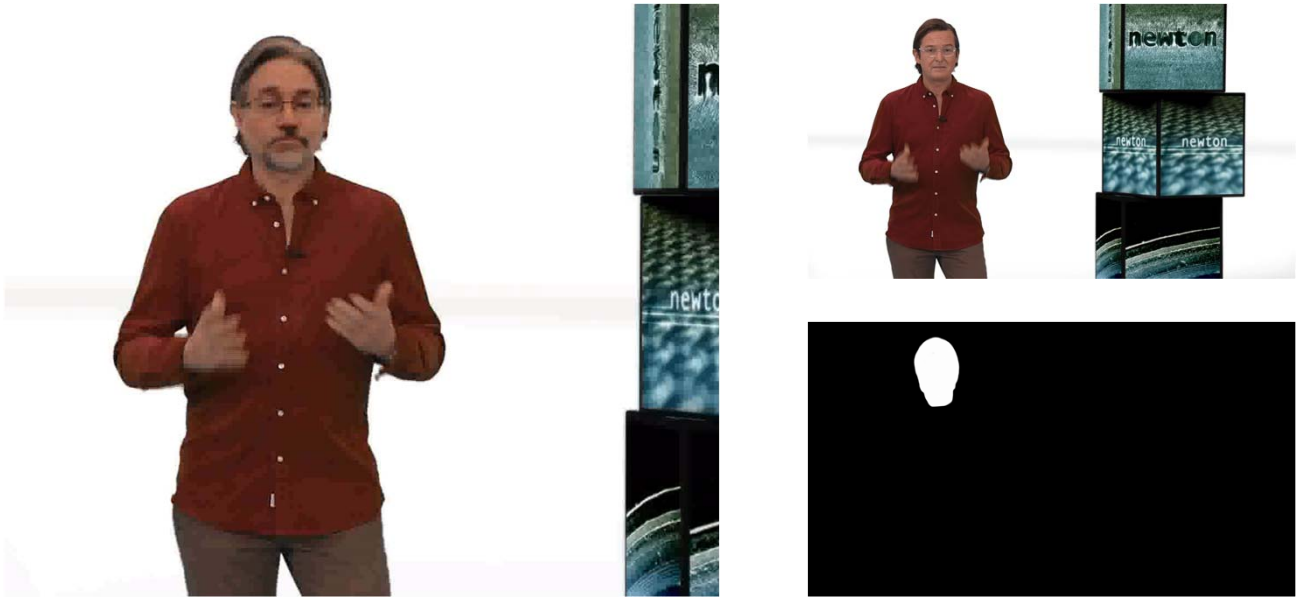


Initial Model Training

After 36 hours of Training

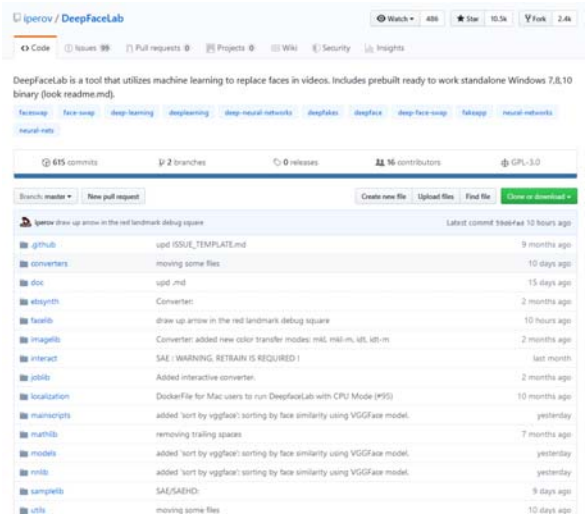


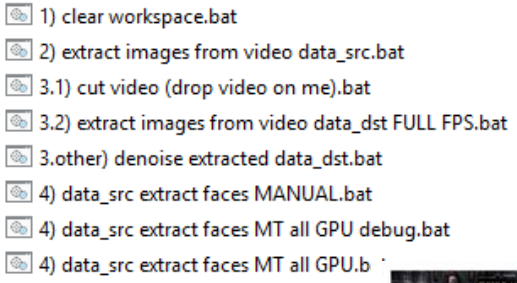
Result



Deep Face Lab

- Available on Github
- Step-wise documentation
- Prepared, enumerated scripts for Windows





AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

NEUE MÖGLICHKEITEN

durch Deep Fakes

De-Aging with Deep Fake

https://www.youtube.com/watch?v=Ddx5B-84ebo

https://www.youtube.com/watch?v=u1OPTCX76Xs

Aging People

https://www.youtube.com/watch?v=3pHJwYLpVCE Mission Impossible Grandpa Version

Indiana Jones

AUSTRIAN INSTITUTE OF TECHNOLOGY

Tom Selleck (original cast)

https://www.youtube.com/watch?v=N2X-GHnijKs

Chris Pratt (test options for reboot)

https://www.youtube.com/watch?v=icGTDi6DU-I

Video Games

Original Videogame Cutscene

https://www.youtube.com/watch?v=26gQVyboD3g

Robert De Niro - Deep Fake edit

https://www.youtube.com/watch?v=D9K7gLHqCOs&t=64s

New Possibilities

- All this took **teams of VFX artists *months*** to achieve
- Now this can be solved in similar quality on **consumer hardware** in **a few days** by **one person**

But there is also misuse…

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

FAKE NEWS AUF DER SPUR

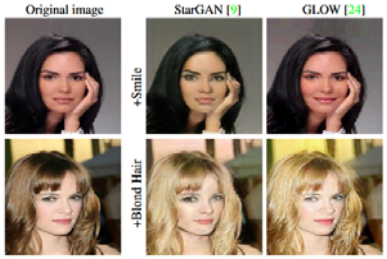
Ein Forschungsschwerpunkt über Medienmanipulation

Types of Fake News

- **Fabricated content**
  - *Completely false*
- **Misleading content**
  - *Misleading use / framing of issue*
- **Imposter content**
  - *Genuine source impersonated with false sources*
- **Manipulated content**
  - *For deception (e.g. images)*
- **False connection**
  - *Headlines, visuals do not support content*
- **False context**
  - *Genuine content shared with false context information*

Fabricated Content

False Context



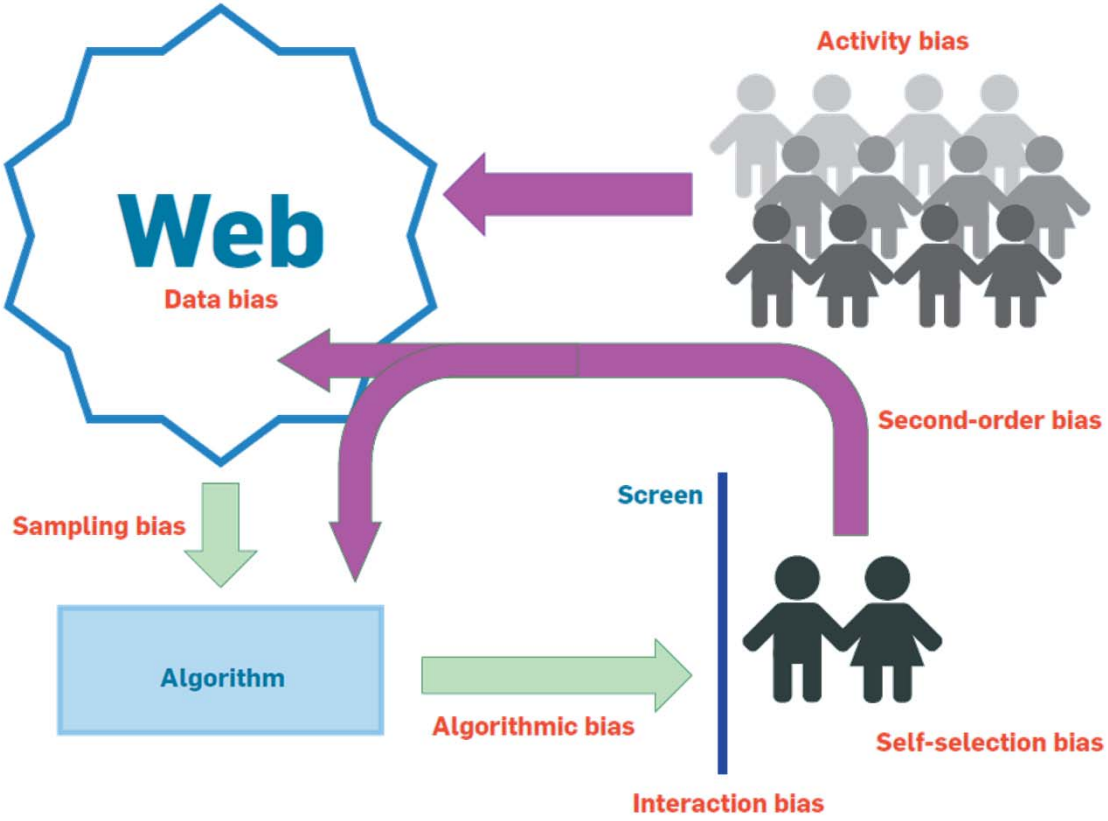
Cozzolino, D., Thies, J., Rössler, A., Riess, C., Nießner, M., & Verdoliva, L. (2018). Forensictransfer: Weakly-supervised domain adaptation for forgery detection. arXiv preprint arXiv:1812.02510.

Manipulated Content

Imposter Content

How Your Audience Will Believe Anything: The Psychology Behind the Fake News https://www.click.co.uk/blog/how-your-audience-will-believe-anything-the-psychology-behind-the-fake-news-bas-van-den-belds-benchmark-2018-talk-review/

The vicious cycle of bias on the Web



Baeza-Yates, Ricardo. "Bias on the web." Communications of the ACM 61.6 (2018): 54-61.

Bedrohungsszenarien

- **Demokratie und staatliche Einrichtungen**
  - Untergrabung des Vertrauens in Parteien, Politiker, Institutionen und Medien
  - Etablierung von Vorstellungen und Ansichten, welche nicht der Realität entsprechen
  - Konkrete Gefährdung von Personen des öffentlichen / politischen Lebens

- **Wirtschaft**
  - Beschädigung der Reputation von Unternehmen
  - Manipulation von Börsenkursen
  - Produktbewertungen
  - Gefährdung durch Falschinformationen im Gesundheitsbereich

Herausforderungen

- Hohe Einsätze und viele Beteiligte
- Böswillige Absicht
- Mangelndes Bewusstsein (Digital Literacy)
- Hohe Ausbreitungsdynamik
- Ständig im Wandel
- Technologisch komplex (unterschiedliche Modalitäten – Ton, Bild, Text, Netzwerk, …)

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

BILD MANIPULATION

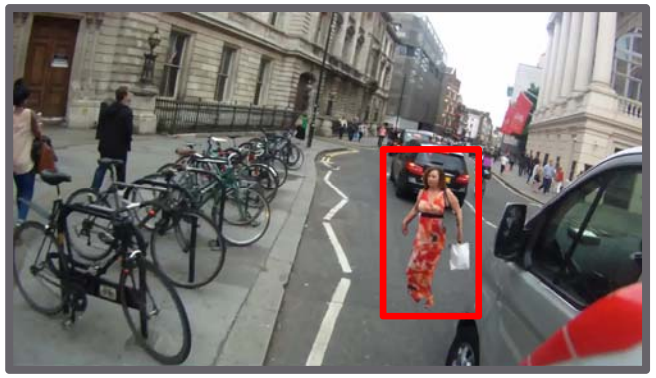
Fotos & Videos

Video MANIPULATION
Schneiden, Kopieren…

NC2016 Datensatz und Ergebnisse (2018):
https://www.nist.gov/itl/iad/mig/nimble-challenge-2017-evaluation

- **Media Forensic Challenge (MFC)**
  https://www.nist.gov/itl/iad/mig/media-forensics-challenge-2019-0
  - ~100.000 manipulierte Bilder
  - ~4000 manipulierte Videos
  - ~5000 annotierte Bilder
  - ~500 annotierte Videos
- **Erkennt Manipulation**

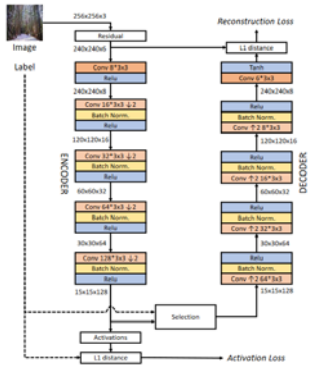
…Nachbearbeitung



- **KI basierte Algorithmen**
- Algorithmen lernen
  - aus großen repräsentativen Trainingsdatensätzen
  - mögliche Artefakten zu erkennen (z.B. starke Kontrastunterschiede, unnatürliche Grenzen, Rauschinkonsistenzen, etc.
- **Erkennen Manipulationen**

Forgery Detection

- Identify GAN generated / forged images
- Encoder-Decoder Network
- Class Activation Mapping (CAM)
- Visualize / explain faked image regions



Cozzolino, D., Thies, J., Rössler, A., Riess, C., Nießner, M., & Verdoliva, L. (2018). Forensictransfer: Weakly-supervised domain adaptation for forgery detection. arXiv preprint arXiv:1812.02510.

Figure 2: Two examples of images manipulated with Face2Face [39] and FaceSwap [2] (left) and their corresponding class activation maps, when the network (XceptionNet [10]) is trained on Face2Face forgeries (middle) and when it is trained on FaceSwap ones (right).

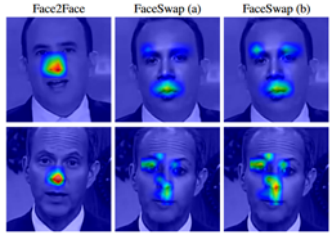
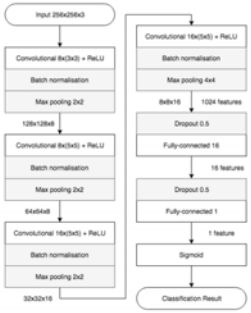
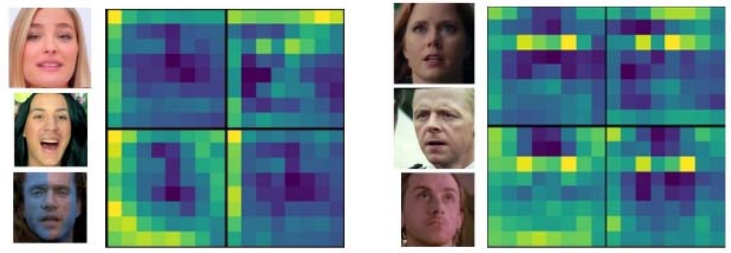


Figure 3: Class activation maps for our method when it is trained on Face2Face [39] and tested on Face2Face forgeries (left) or FaceSwap [2] (middle), and finally trained on Face2Face but fine-tuned using only four images manipulated with FaceSwap and tested on FaceSwap (right).

Detecting GAN generated Faces

- Identify GAN generated faces
- Train on DeepFake Dataset
- High accuracy with simple approach
- Observation: Activations
  - Fake ➔ Background
  - Real ➔ Eyes



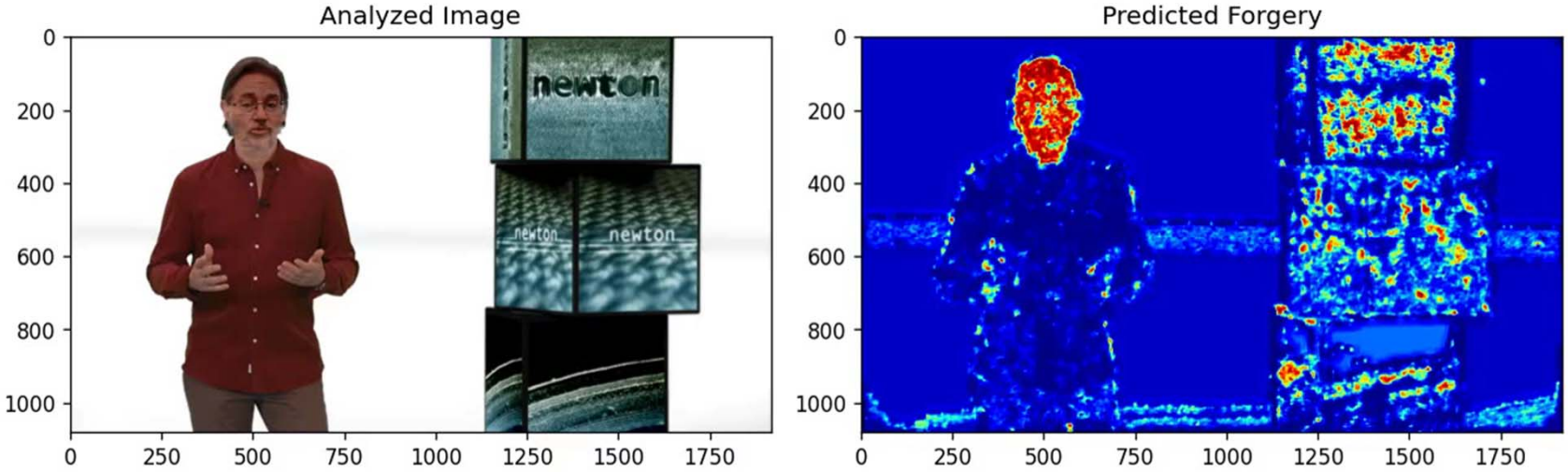


mean layer output of 100 *deepfake* faces

mean layer output of 100 real faces

Marra, F., Gragnaniello, D., Cozzolino, D., & Verdoliva, L. (2018, April). Detection of GAN-generated fake images over social networks. In 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) (pp. 384-389). IEEE.

Analysing our Deep Fake



AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

AUDIO MANIPULATION

Sprache & Videos

Tacotron-2 Text-to-Speech

- Generate Speech from Text

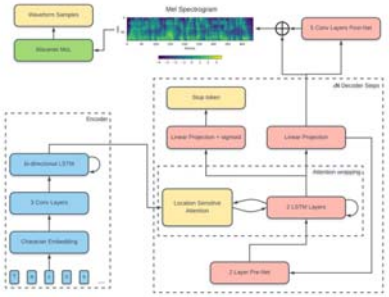
"That girl did a video about Star Wars lipstick."

"She earned a doctorate in sociology at Columbia University."

"George Washington was the first President of the United States."

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., ... & Saurous, R. A. (2018, April). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4779-4783). IEEE.

Model Architecture:



- Stress and intonation
- Questions
- Prosody
  - Intonation, rhythm, tone

Style / Reference

Reference text: Alice was not much surprised at this, she was getting so used to queer things happening.

Base Voice Model

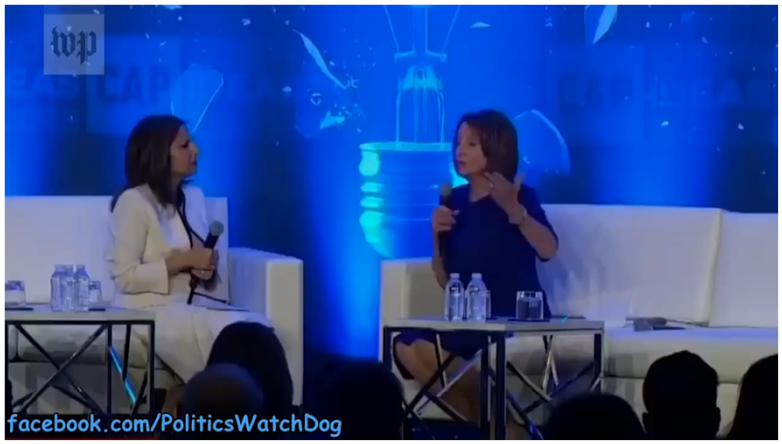
Result

Perturbed text: Eric was not much surprised at this, he was getting so used to TensorFlow breaking.

Singing

Skerry-Ryan, R. J., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., ... & Saurous, R. A. (2018). Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. arXiv preprint arXiv:1803.09047.

Simple but effective / No AI



Pelosi videos manipulated to make her appear drunk are being shared on social media
https://www.youtube.com/watch?v=sDOo5nDJwgA

FFmpeg

Suche

Anmelden | Einstellungen | Hilfe/Anleitung | Über Trac | Register

Wiki | Journal | Tickets anzeigen | Suche | Tags

Wiki: How to speed up / slow down a video

+1 | Aufwärts | Startseite | Inhaltsverzeichnis | Änderungshistorie

Speeding up/slowing down video

You can change the speed of a video stream using the setpts video filter. Note that in the following examples, the audio stream is not changed, so it should ideally be disabled with `-an`.

To double the speed of the video, you can use:

```
ffmpeg -i input.mkv -filter:v "setpts=0.5*PTS" output.mkv
```

The filter works by changing the presentation timestamp (PTS) of each video frame. For example, if there are two succesive frames shown at timestamps 1 and 2, and you want to speed up the video, those timestamps need to become 0.5 and 1, respectively. Thus, we have to multiply them by 0.5.

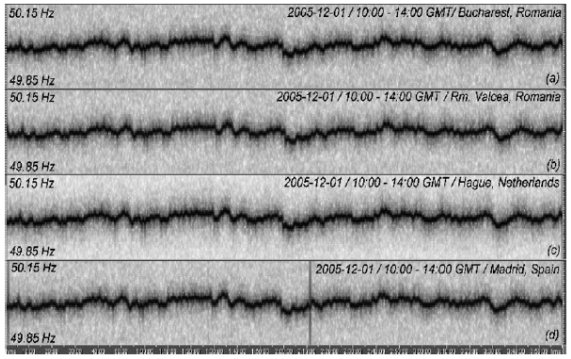
Note that this method will drop frames to achieve the desired speed. You can avoid dropped frames by specifying a higher output frame rate than the input. For example, to go from an input of 4 FPS to one that is sped up to 4x that (16 FPS):

```
ffmpeg -i input.mkv -r 16 -filter:v "setpts=0.25*PTS" output.mkv
```

To slow down your video, you have to use a multiplier greater than 1:

```
ffmpeg -i input.mkv -filter:v "setpts=2.0*PTS" output.mkv
```

AUDIO MANIPULATION (1/2)
Schneiden, Kopieren/Verschieben, Einfügen



Grigoras, C. (2007). Applications of ENF criterion in forensic audio, video, computer and telecommunication analysis. Forensic science international, 167(2-3), 136-145.

- **Electronic Network Frequency (ENF)**
- Eindeutiges Muster: Frequenz des Stromnetzes in der Audio-Aufnahme
- Batterie?, Nachbearbeitung?
- **Erkennt Manipulation**
  + Ort/Zeit der Aufnahme

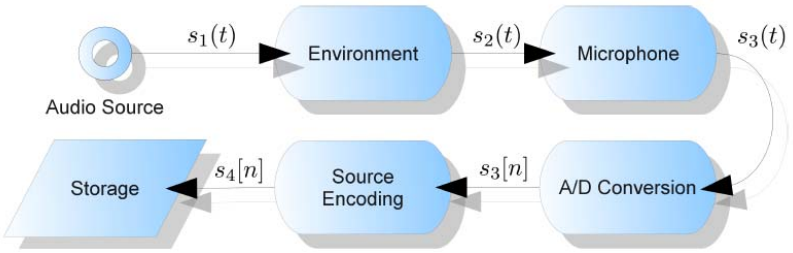


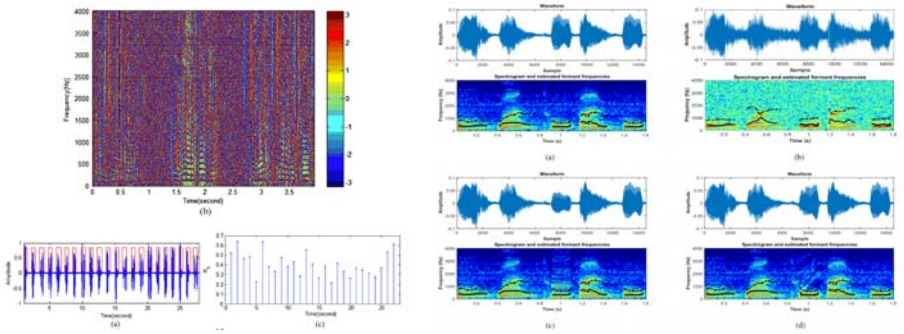
Fig. 1. Mobile Recording - Process Flow

Cuccovillo, L., Mann, S., Tagliasacchi, M., & Aichroth, P. (2013, September). Audio tampering detection via microphone classification. In 2013 IEEE 15th International Workshop on Multimedia Signal Processing (MMSP) (pp. 177-182). IEEE.

- **Mikrofon Klassifizierung**
- Mikrofon und A/D HW-Komponenten haben Einfluss auf die Aufnahme
- Robust gegenüber Komprimierungsverlusten
- **Erkennt Manipulation**

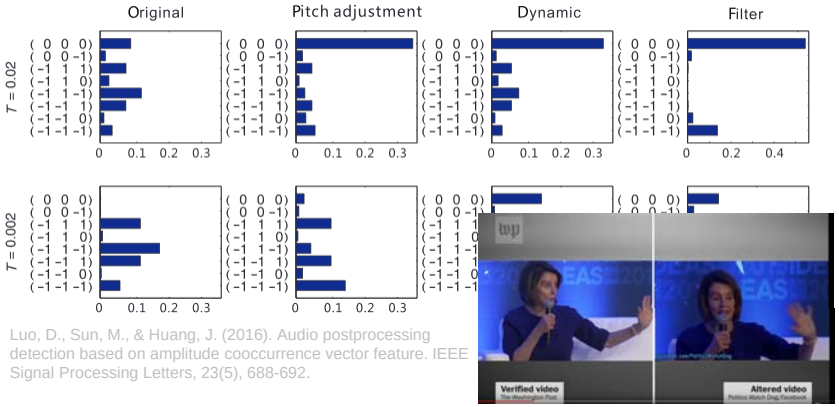
AUDIO MANIPULATION (2/2)

Schneiden, Kopieren…



- **Phasen/Ton Analyse**
- Manipulation zerstört Phasenverteilung
- Klangbild-Analyse robust zu Manipulation z.B. der Tonhöhe
- **Erkennt Manipulation + Identifiziert Quellen**

…Nachbearbeitung



Luo, D., Sun, M., & Huang, J. (2016). Audio postprocessing detection based on amplitude cooccurrence vector feature. IEEE Signal Processing Letters, 23(5), 688-692.

- **Amplitude Cooccurrence Vectors (ACV)**
- Textur-Analyse im Frequenzbereich
- Manipulation ändert Wahrscheinlichkeit für gemeinsames Vorkommen von Audio Samples
- **Erkennt Manipulation + Art der Manipulation** (Tonhöhe, Filter, Geschwindigkeit)

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

TEXT ANALYSE

Desinformation in Text

CONTENT-BASED APPRoAChES

- Headline & body text

- **Linguistic cues and patterns**
  - Character, word, sentence or document level
  - Linguistic Inquiry and Word Count (LIWC)

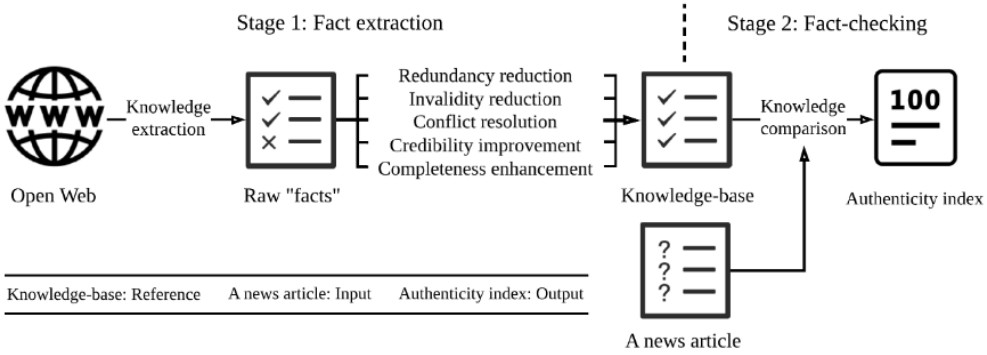
- **Style-based**
  - Hashtags, mentions, punctuation marks, sentiment
  - Topics languages, domains

Attribute Type	Feature
Quantity	Character count
	Word count
	Noun count
	Verb count
	Number of noun phrases
	Sentence count
	Paragraph count
	Number of modifiers (e.g., adjectives and adverbs)
Complexity	Average number of clauses per sentence
	Average number of words per sentence
	Average number of characters per word
	Average number of punctuations per sentence
Uncertainty	Percentage of modal verbs
	Percentage of centainty terms
	Percentage of generalizing terms
	Percentage of tentative terms
	Percentage of numbers and quantifiers
	Number of question marks
Subjectivity	Percentage of subjective verbs
	Percentage of report verbs
	Percentage of factive verbs
	Percentage of imperative commands

Knowledge-based approaches

- Check for truthfulness of claims

- **Manual fact-checking**
  - Expert-based
  - Crowdsources
  - politifact.com, snopes.com



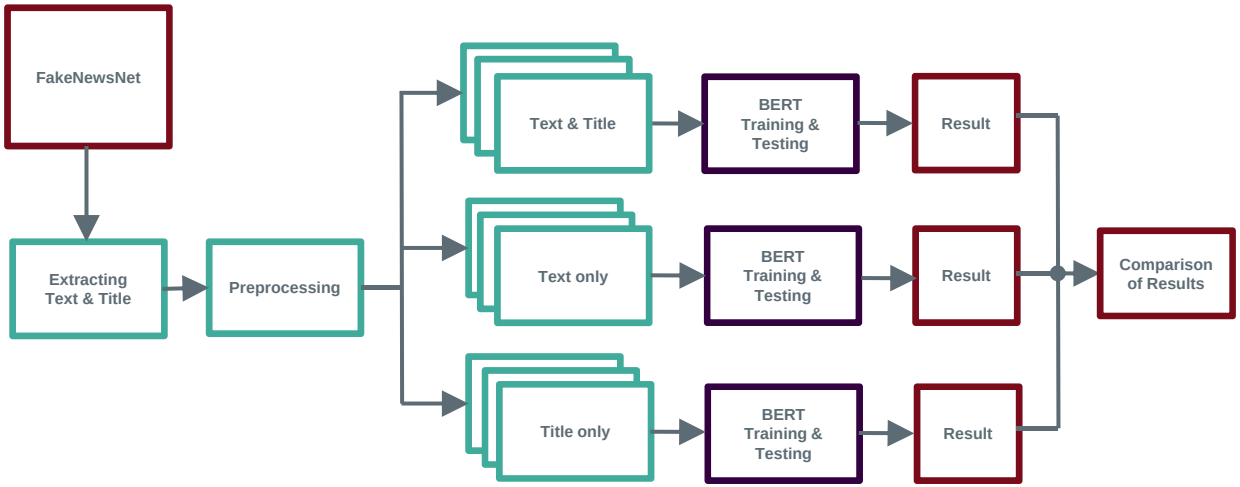
- **Automatic fact-checking**
  - Linked Open Data (i.e. DBpedia)
  - Fact-extraction (Knowledge base construction)
  - Fact-checking (Knowledge comparison

SOCIAL-CONTEXT Based APPROACHES

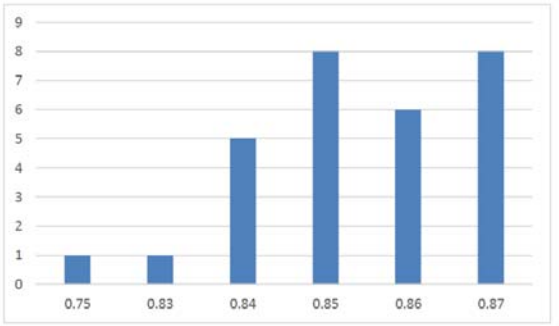
- **Stance-based**
  - Stance of body text relative to the headline claim
  - Viewpoint of user
  - Infer validity of original article
  - Support or refute claim
- **User-based**
  - Registration age
  - Numer of followers / followees
- **Propagation-based**
  - Propagation networks
  - i.e. Twitter shares / retweets
  - Likes
- **Credibility-based**
  - Headlines (clickbaits)
  - News source, spreaders, author

Experiments – Fake News Detection

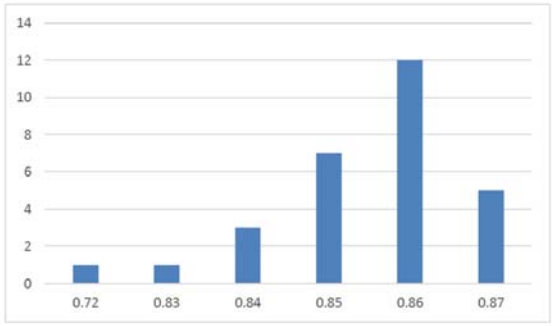
Masterarbeit: Mina Schütz, *Detection and identification of fake news - Binary Content Classification with a Pre-Trained Language Model*



Validation – Average: 85.18%

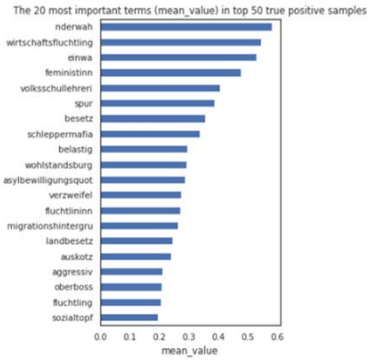


Testing – Average: 85.14%

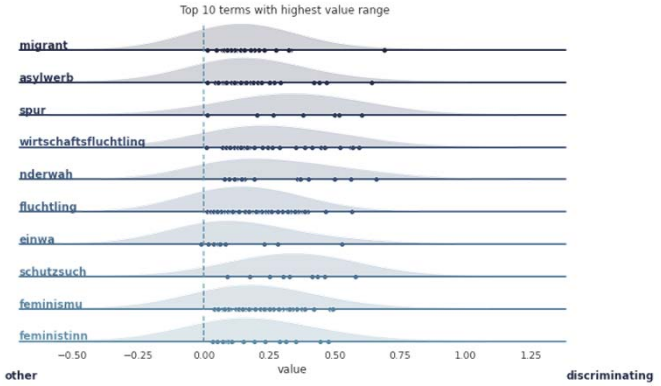


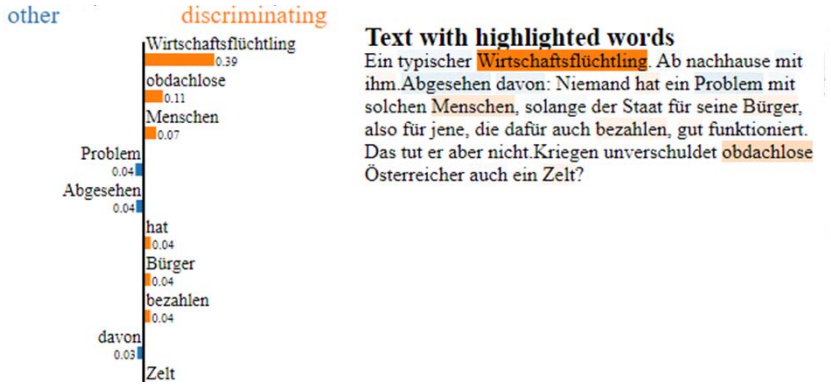
Explaining Hate Speech Models

Masterarbeit: Daria Liakhovets.



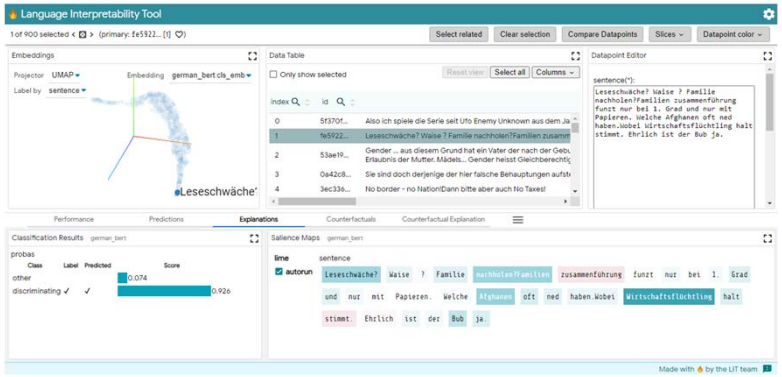
True Positives





Text with highlighted words

Ein typischer Wirtschaftsflüchtling. Ab nachhause mit ihm.Abgesehen davon: Niemand hat ein Problem mit solchen Menschen, solange der Staat für seine Bürger, also für jene, die dafür auch bezahlen, gut funktioniert. Das tut er aber nicht.Kriegen unverschuldet obdachlose Österreicher auch ein Zelt?



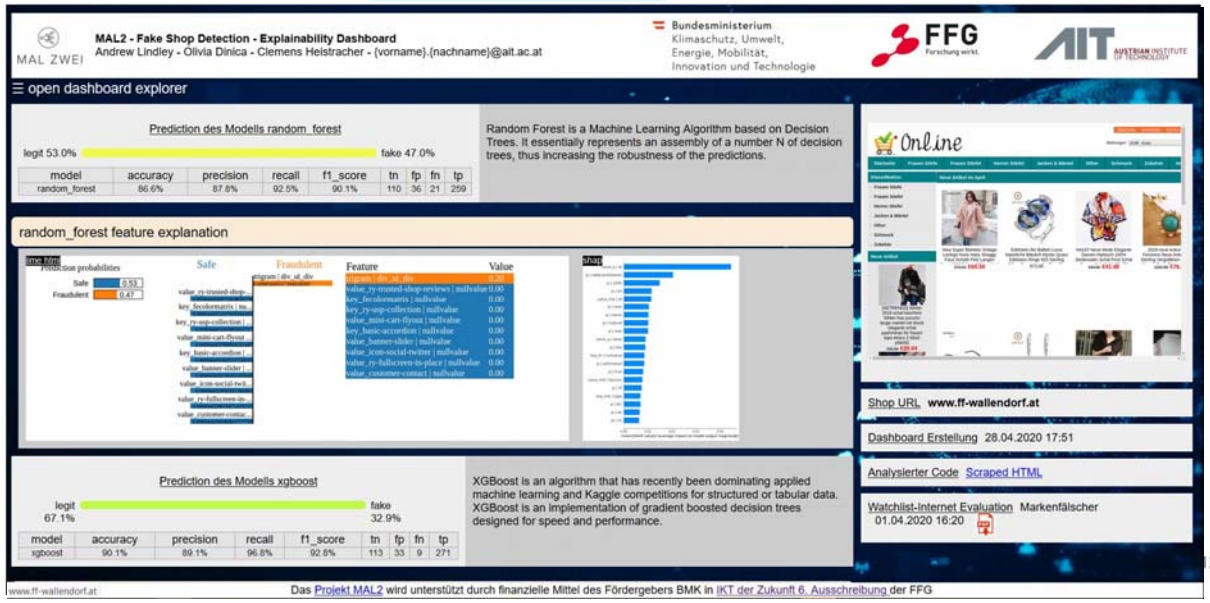
AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

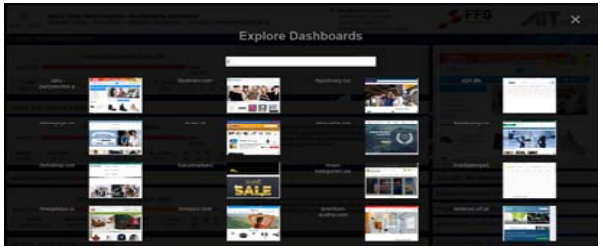
ZUSAMMENFASSUNG ANALYSEVERFAHREN

Querschnitt – Video / Audio

AUSTRIAN INSTITUTE OF TECHNOLOGY

INTERPRETABLE | EXPLAINABLE | TRANSPARENT





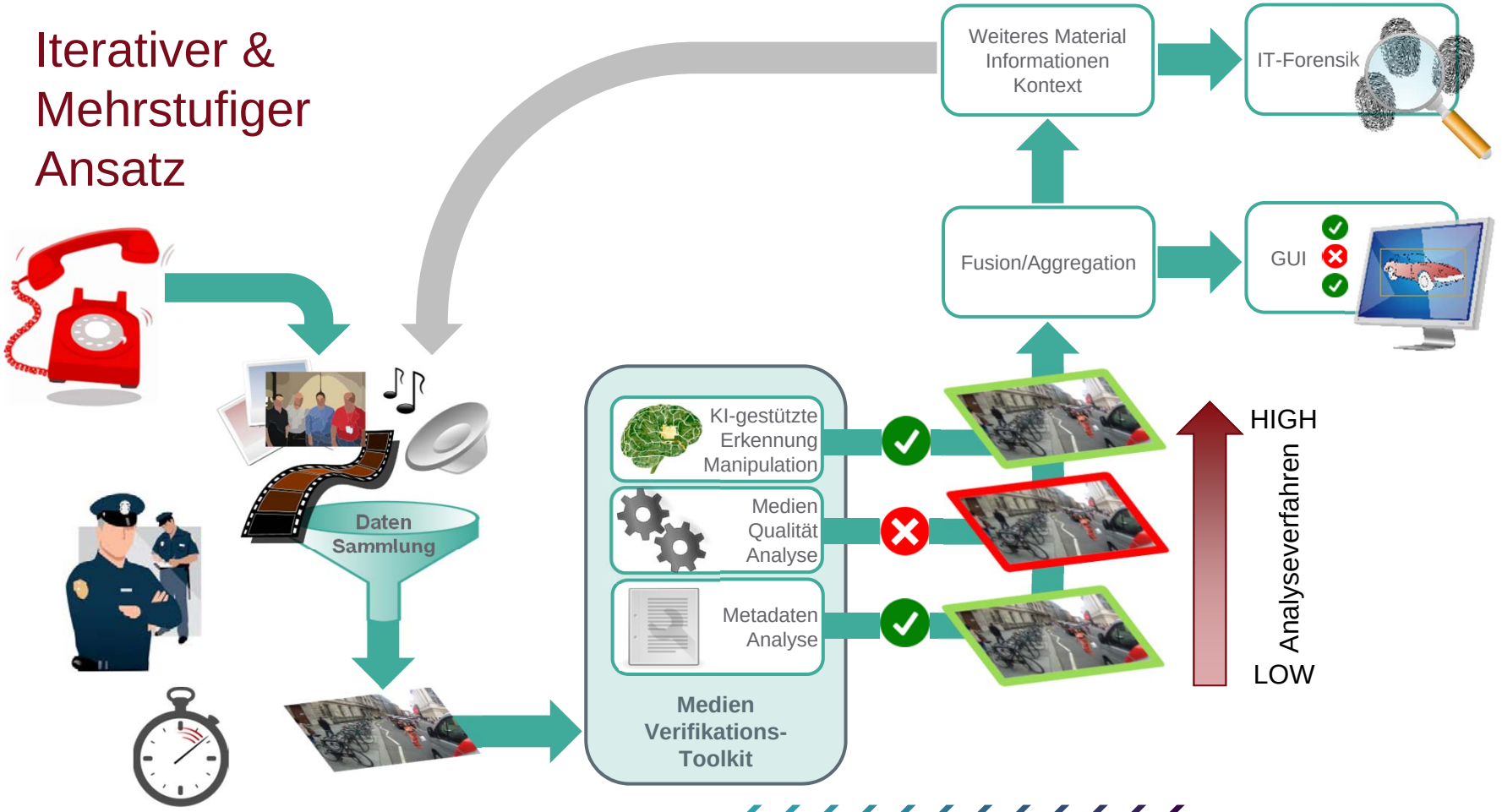

```
Dashboard builder for eCommerce Site to verify

optional arguments:
  -h, --help            show this help message and exit
  -u URL, --url URL     URL of site you want the base url to be verified.
  -i INPUT_DIR, --input-dir INPUT_DIR
                        Relative path to location hosting models and dict
  -o OUTPUT_DIR, --output-dir OUTPUT_DIR
                        Relative path for dasbhoard results
  -f {lime,shap} [{lime,shap} ...], --feature-importance {lime,shap} [{lime,shap} ...]
                        Options are: lime, shap or no flag for None
  --use-cache           Set flag if you want to use locally cached version of
                        scraped html files - re-scrapes only when site not
                        available locally
  --submit-results      Set flag if you want to submit the results of the
                        model prediction to the fake-shop database for manual
                        inspection
  --identify-logos      Set flag if you want to scrape images and apply the
                        KOSOH trustmark/payment-provider image identifier
```

Erklärbarkeit und ExpertInnen Tools

Das Fake-Shop Expert-Analysis Dashboard ist eine Komponente, die für die Zielgruppe der FachexpertInnen der Watchlist Internet umgesetzt wurde, um mit den Fake-Shop Bewertungsmodellen in der täglichen Praxisanwendung zu interagieren und beliebige Fake-Shop Verdachtsmomente gegenüber der Vorhersage der trainierten Modelle detailliert zu validieren. (als virtuelle Maschine verfügbar)

Iterativer & Mehrstufiger Ansatz



AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

Thank you!

Alexander Schindler

16.10.2020