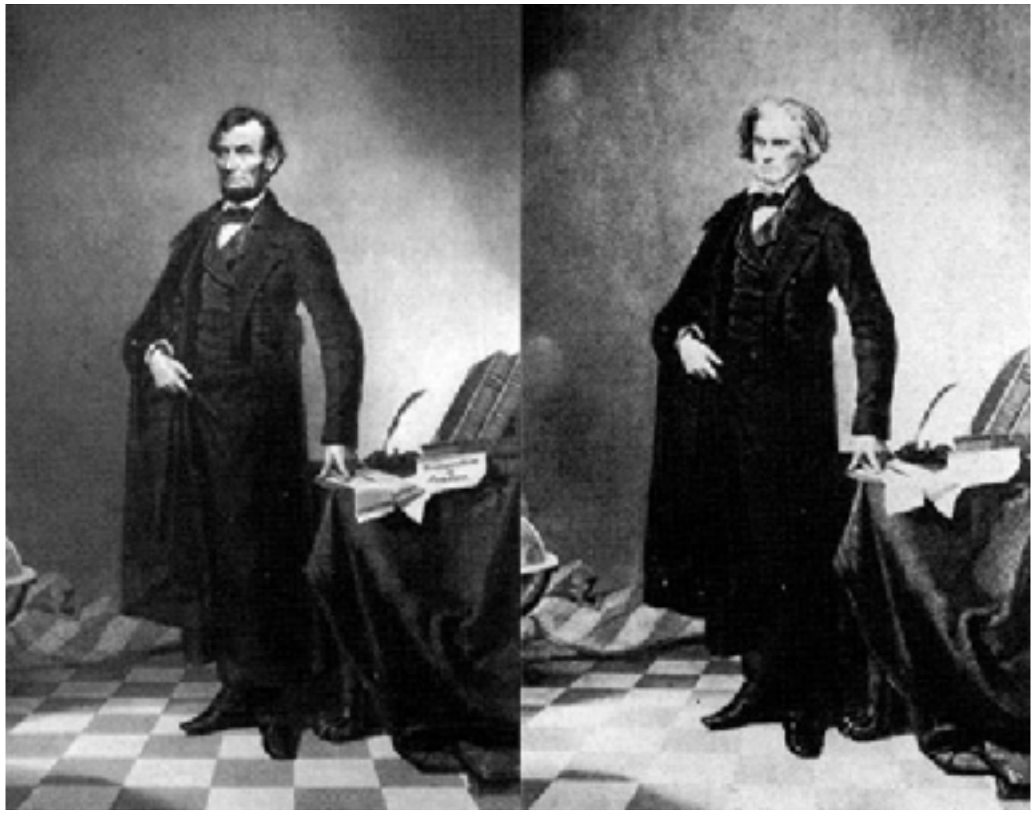
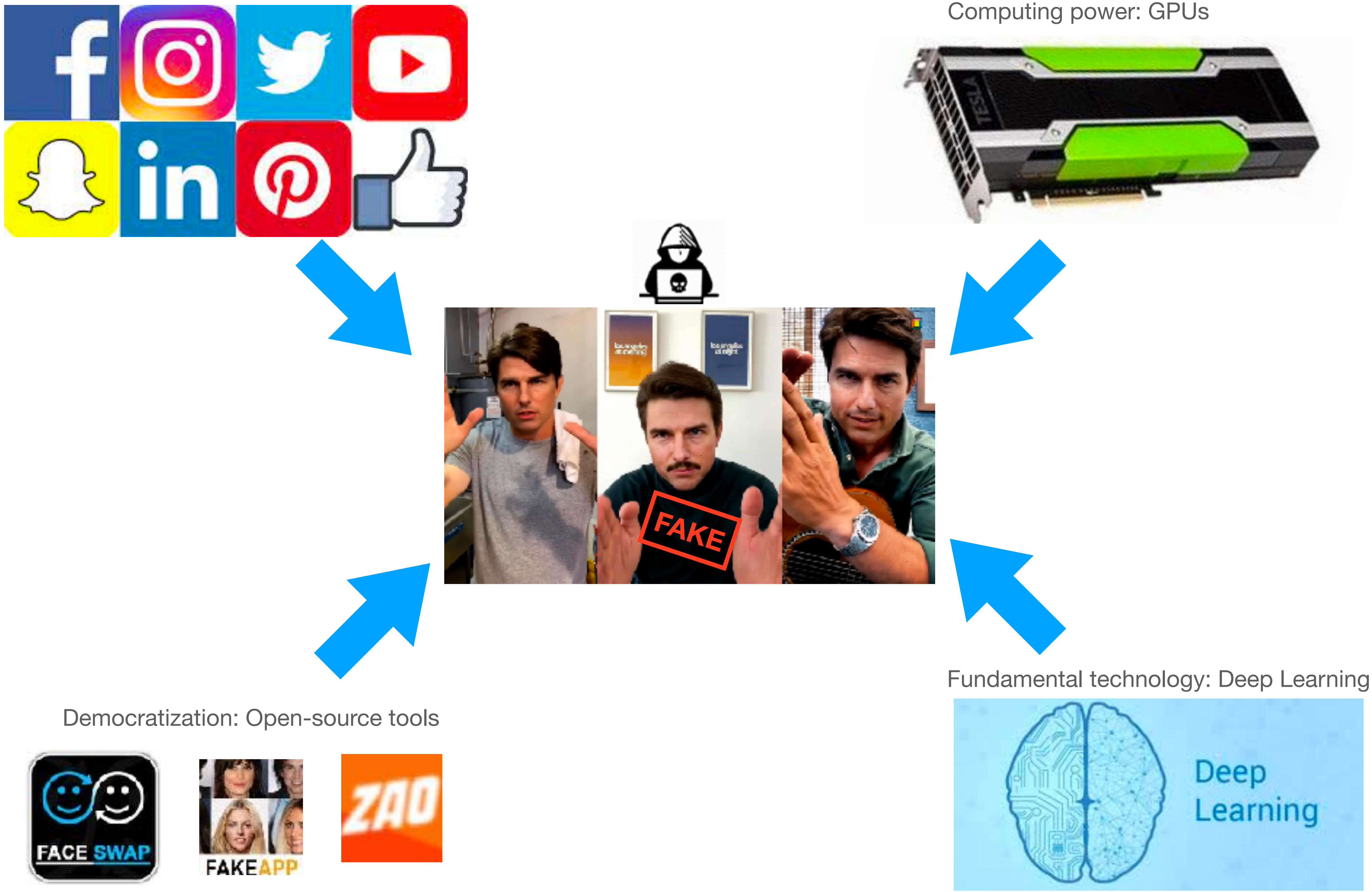
Fighting AI Synthesized Fake Media

Siwei Lyu
Department of Computer Science
and Engineering
University at Buffalo,
State University of New York
Email: siweilyu@buffalo.edu

Fake media are not new …



Deep(learning)Fake media are recent



Data source and dissemination channel: social platform

Computing power: GPUs

Democratization: Open-source tools

Fundamental technology: Deep Learning

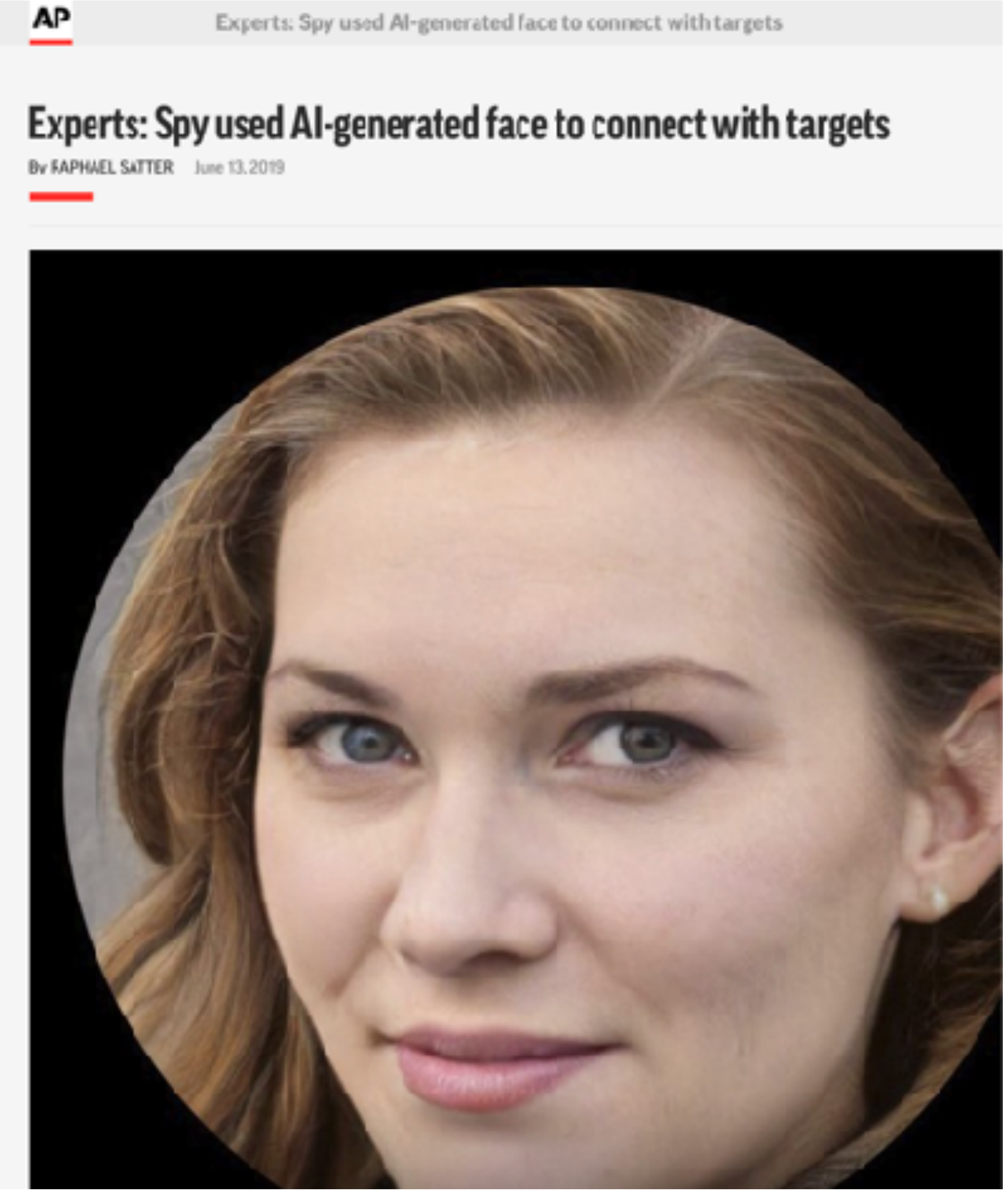
GAN faces

We see faces …

Image source: www.thispersondoesnotexist.com

of persons who do not exist;

GAN faces

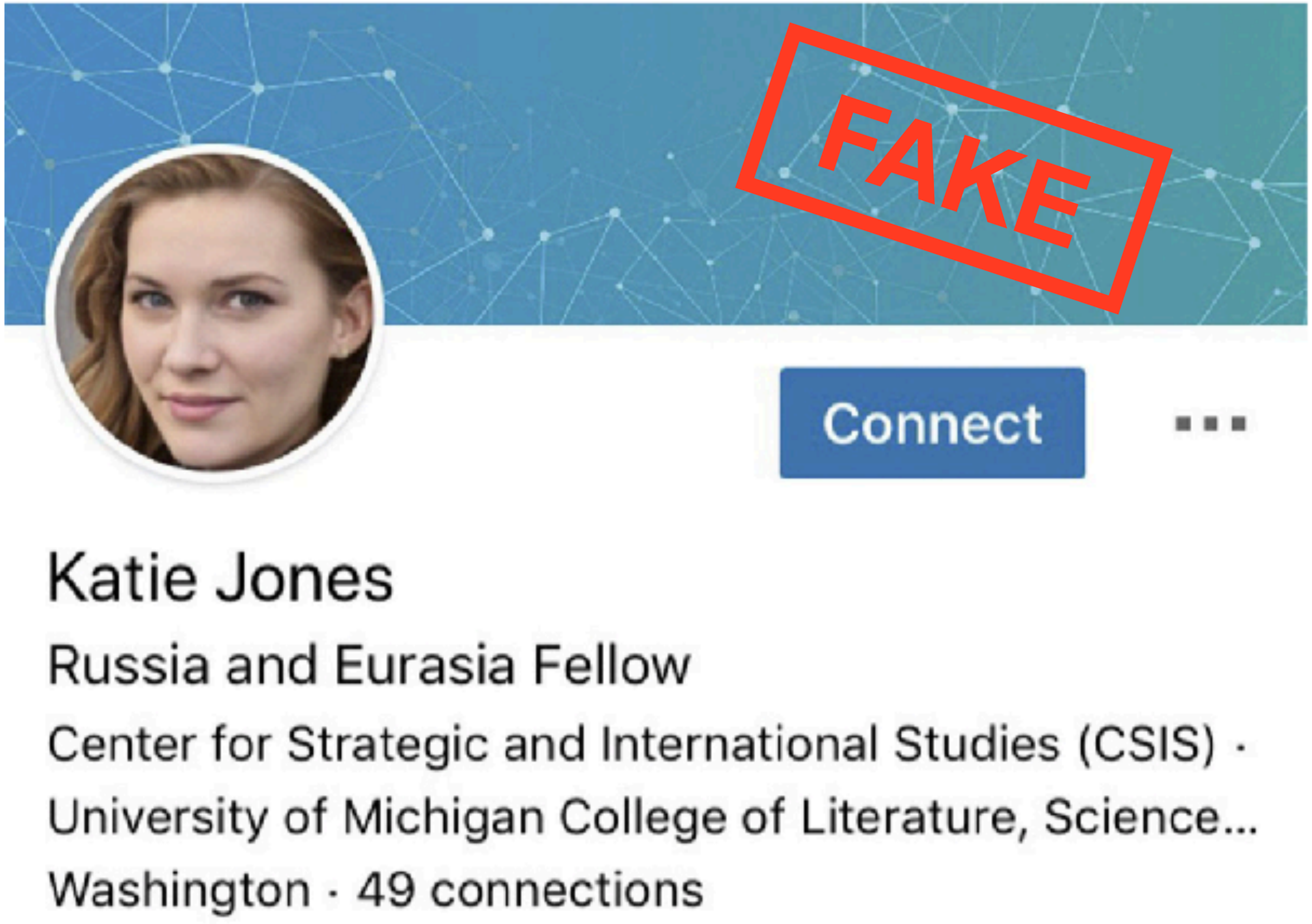


AP

Experts: Spy used AI-generated face to connect with targets

Experts: Spy used AI-generated face to connect with targets

By RAPHAEL SATTER June 13, 2019



Katie Jones

Russia and Eurasia Fellow

Center for Strategic and International Studies (CSIS) ·

University of Michigan College of Literature, Science...

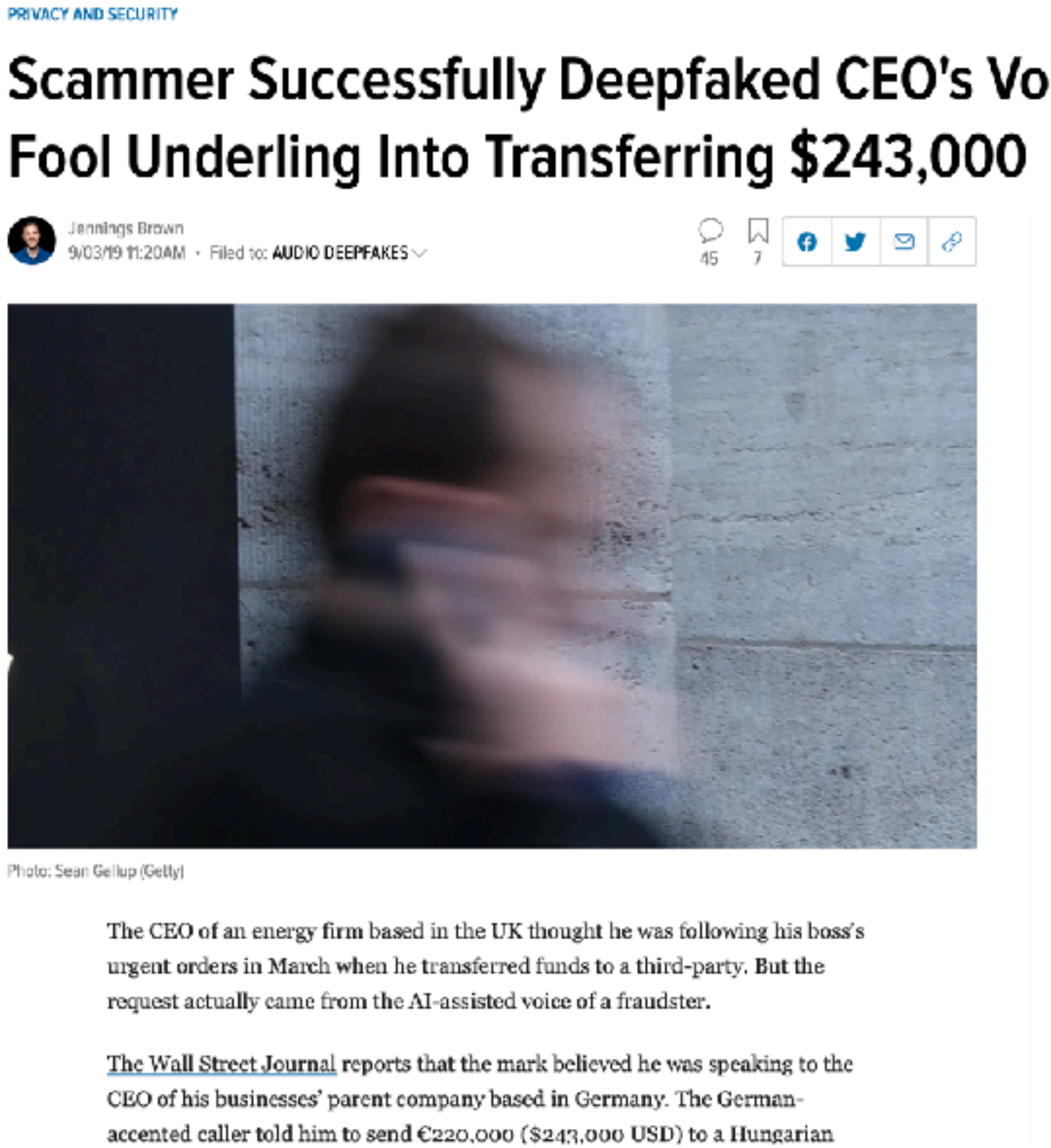
Washington · 49 connections

Audio DeepFakes

We hear speeches …

that were not spoken;

Audio DeepFakes



PRIVACY AND SECURITY

Scammer Successfully Deepfaked CEO's Voice To Fool Underling Into Transferring $243,000

Jennings Brown
9/03/19 11:20AM · Filed to: AUDIO DEEPFAKES

Photo: Sean Gallup (Getty)

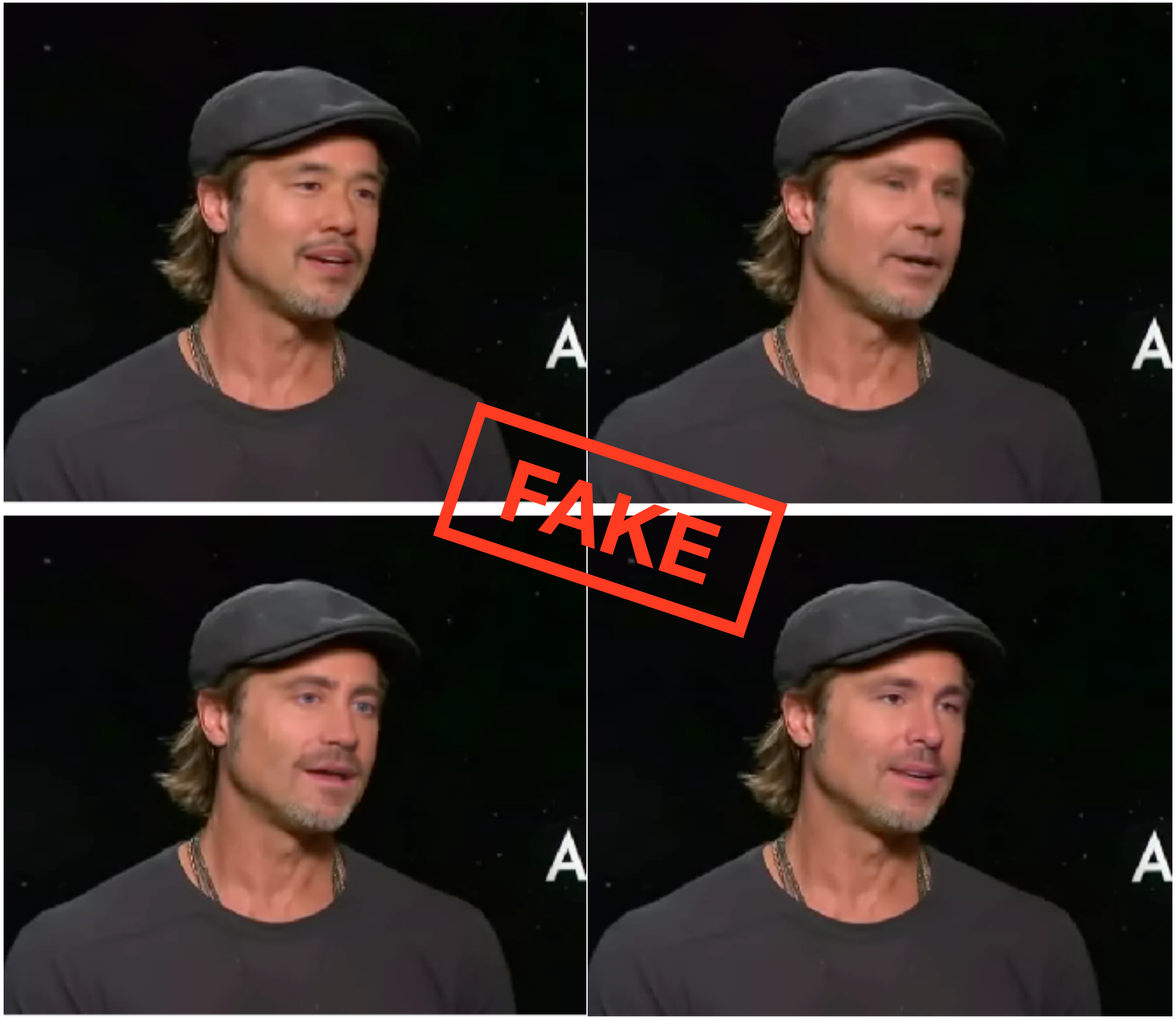
The CEO of an energy firm based in the UK thought he was following his boss's urgent orders in March when he transferred funds to a third-party. But the request actually came from the AI-assisted voice of a fraudster.

The Wall Street Journal reports that the mark believed he was speaking to the CEO of his businesses' parent company based in Germany. The German-accented caller told him to send €220,000 ($243,000 USD) to a Hungarian

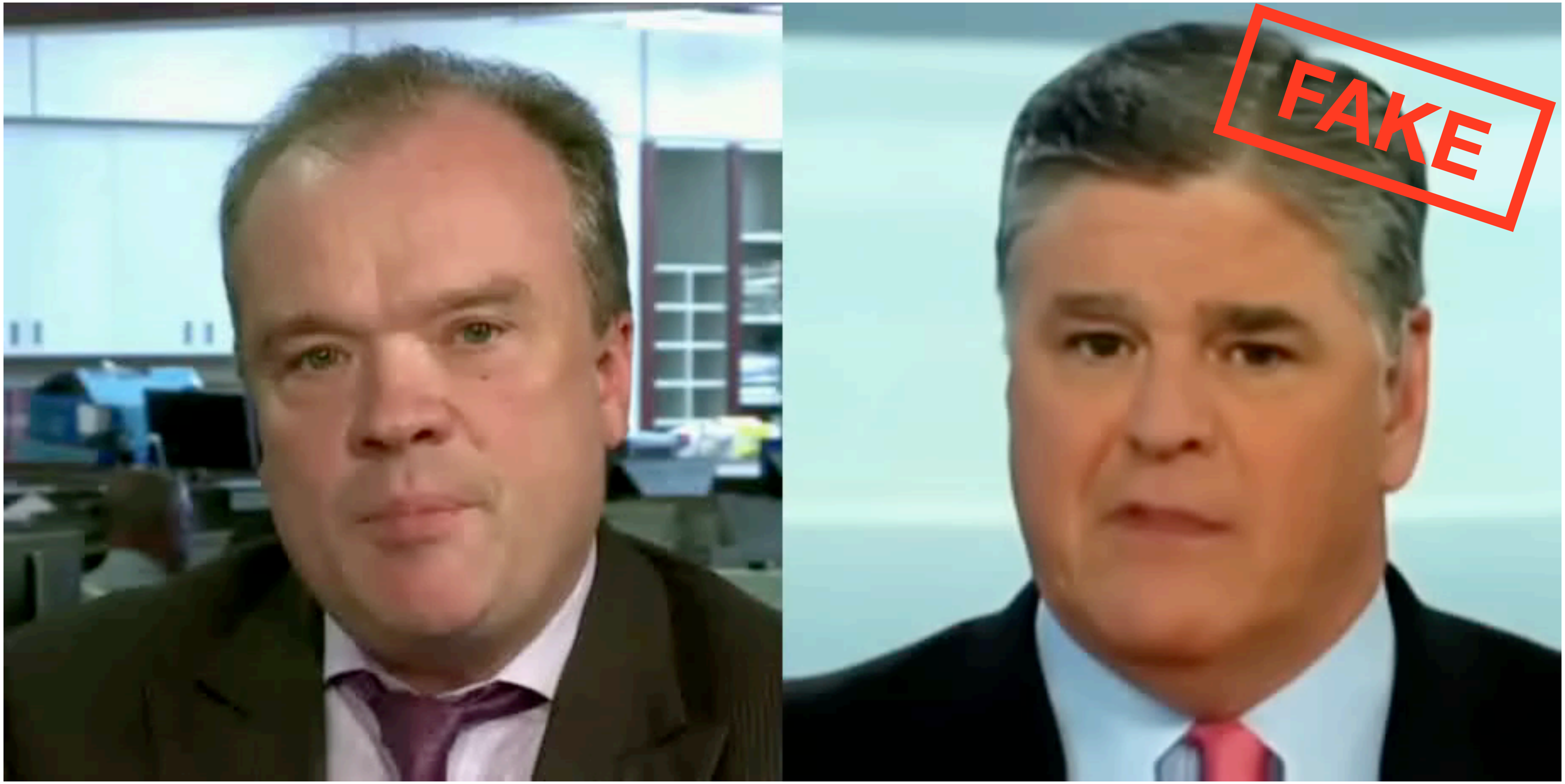
Face-swap videos

And we watch videos …

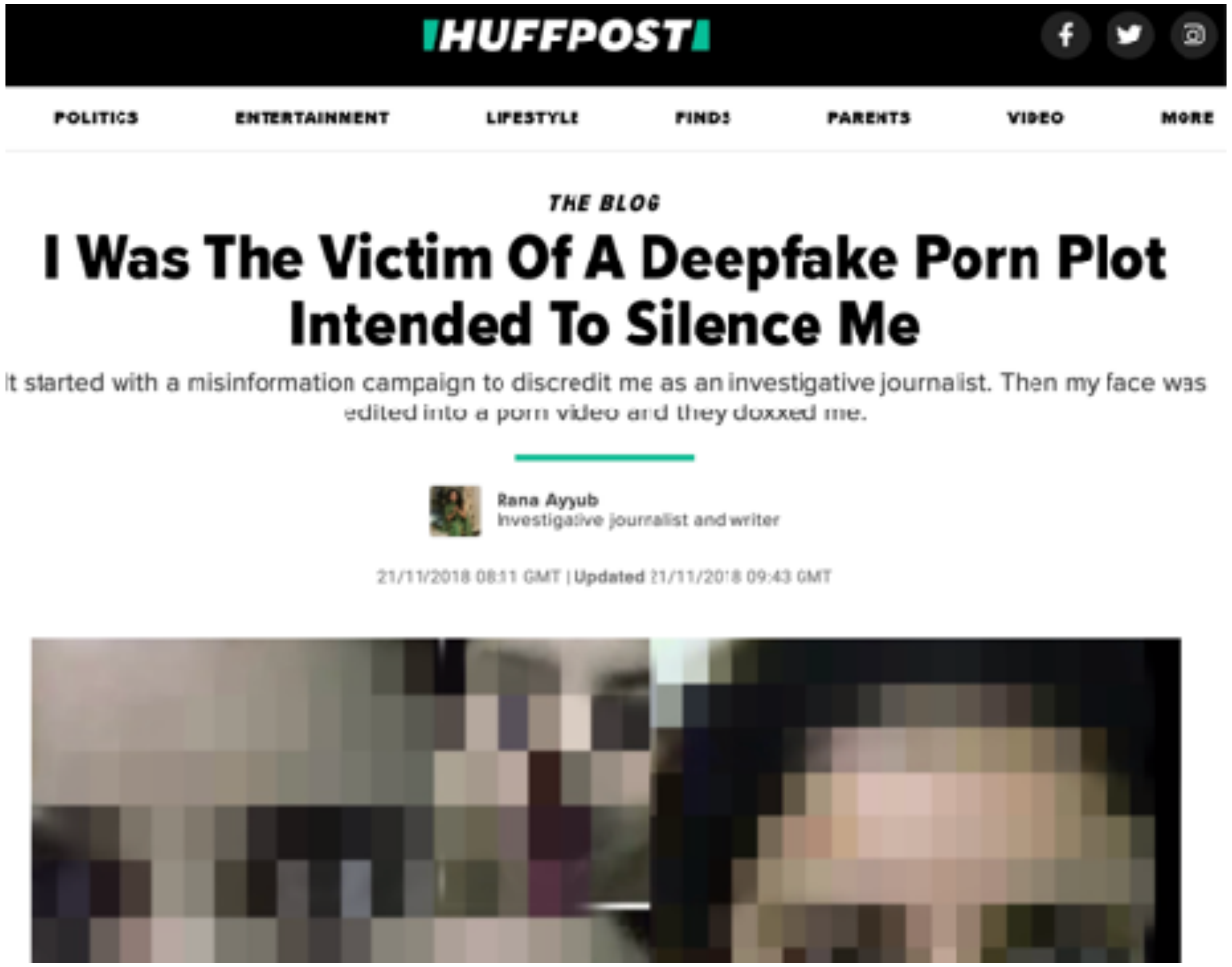
with swapped faces



Face puppetry



Face-swap videos



THE BLOG

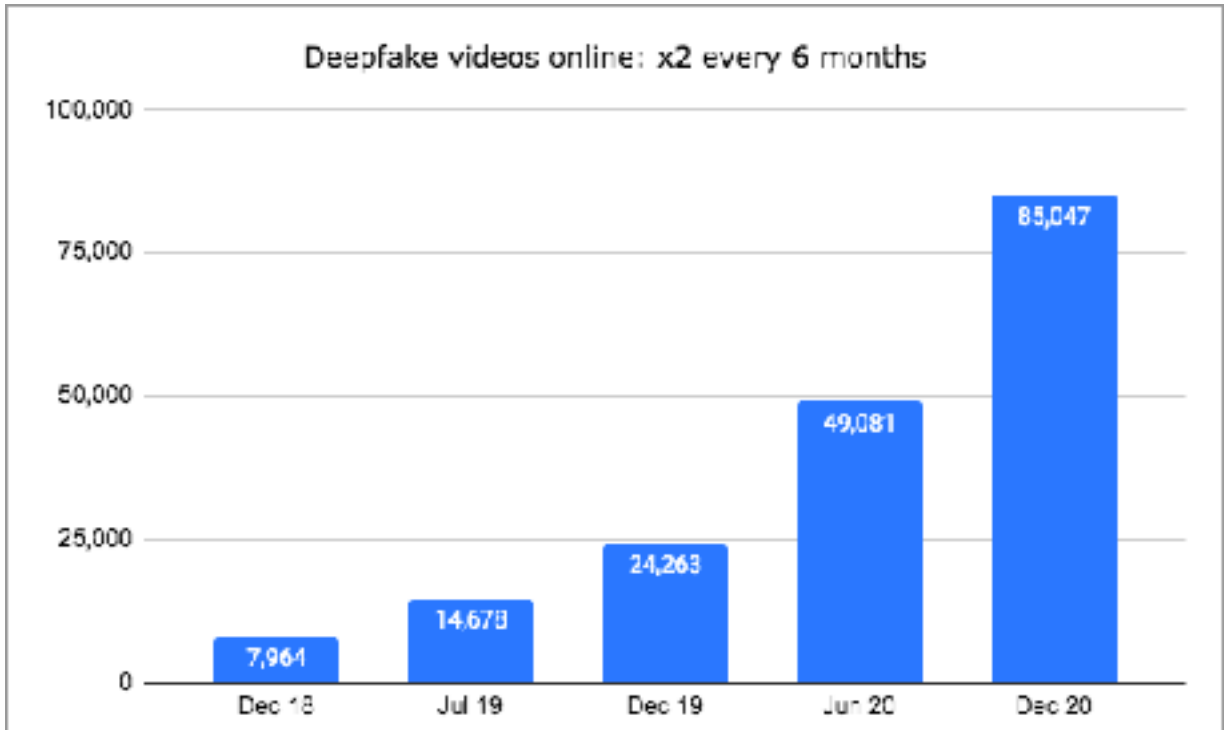
I Was The Victim Of A Deepfake Porn Plot Intended To Silence Me

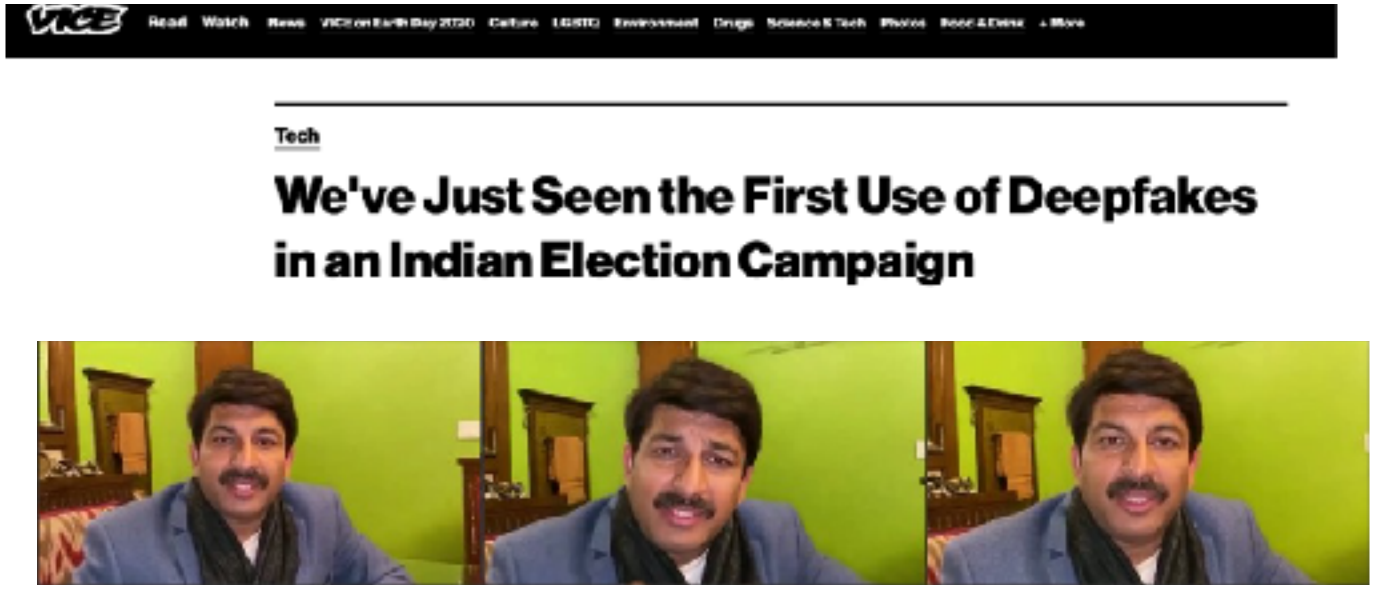
It started with a misinformation campaign to discredit me as an investigative journalist. Then my face was edited into a porn video and they doxxed me.

Rana Ayyub
Investigative journalist and writer

21/11/2018 08:11 GMT | Updated 21/11/2018 09:43 GMT

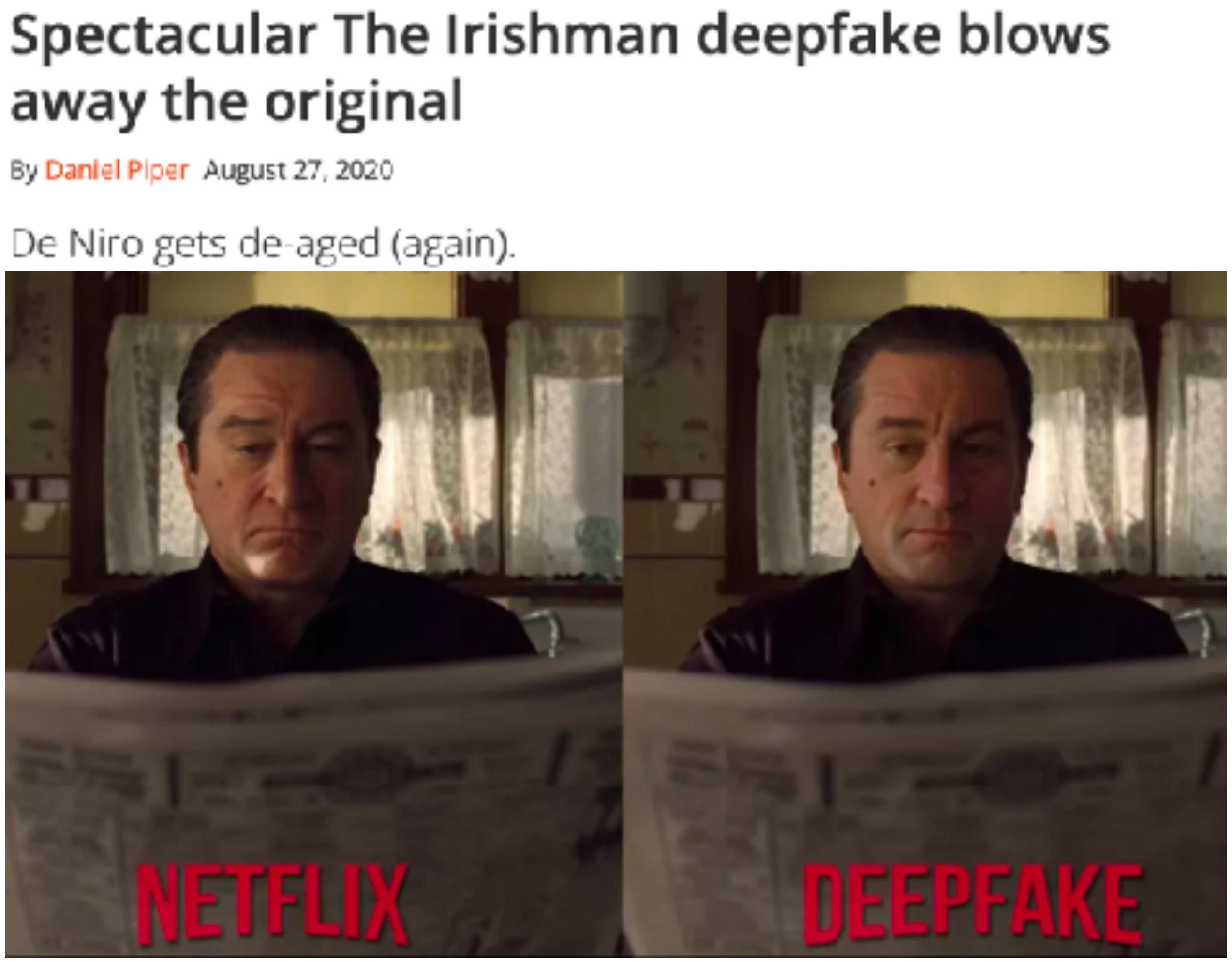
Source: https://sensity.ai/





Tech

We've Just Seen the First Use of Deepfakes in an Indian Election Campaign

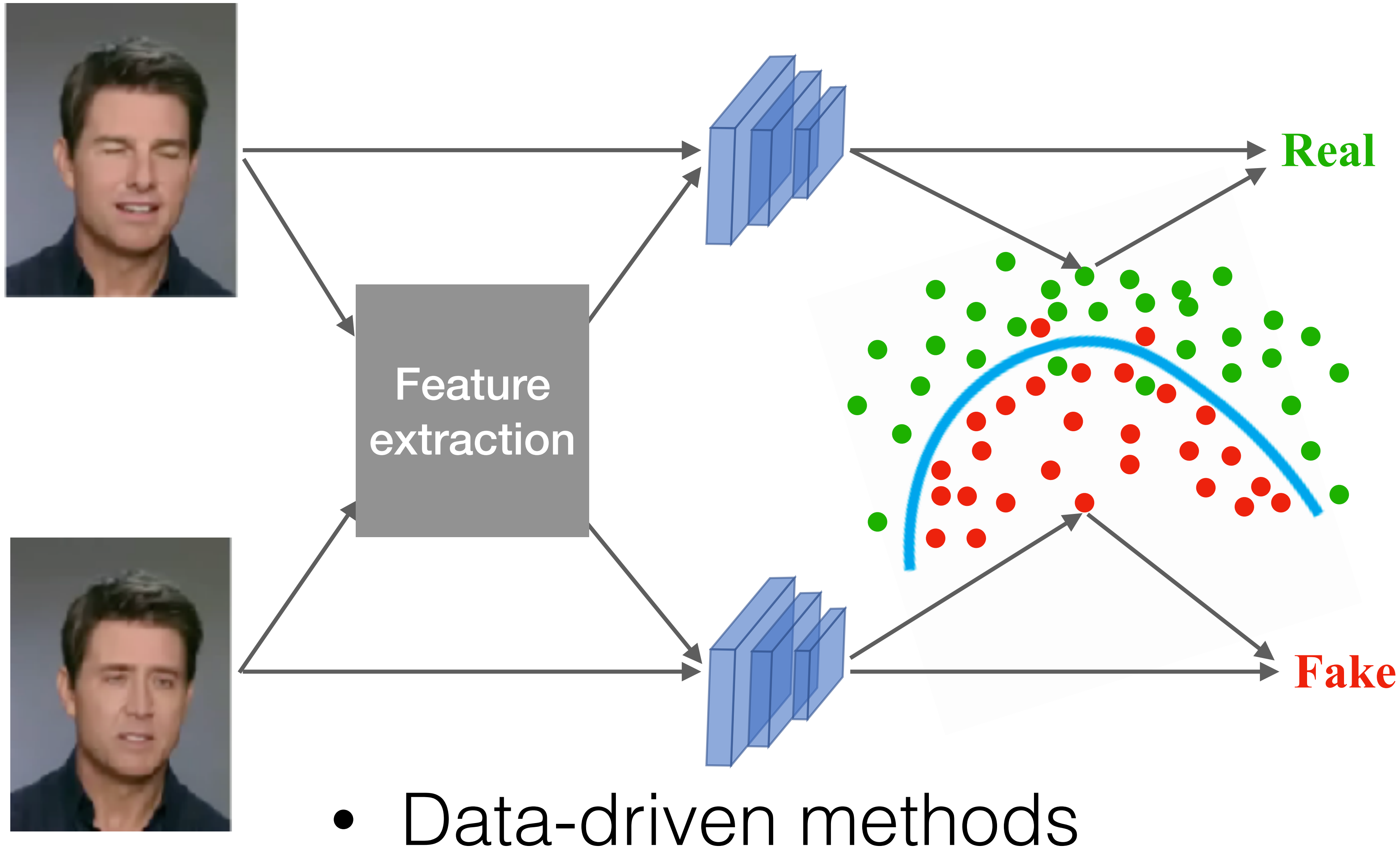


Spectacular The Irishman deepfake blows away the original

By Daniel Piper August 27, 2020

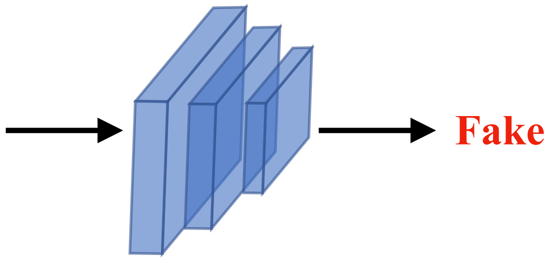
De Niro gets de-aged (again).

DeepFake detection



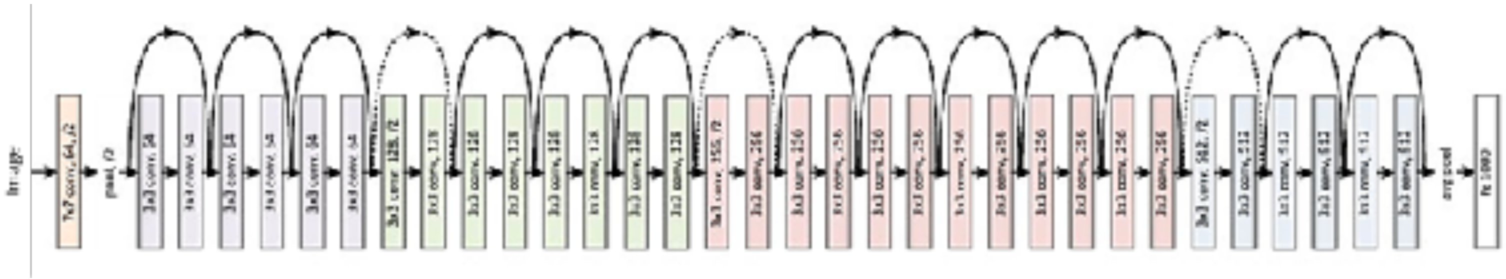
- Data-driven methods
- Cue-based methods
  - Signal features
  - Physical/physiological cues

Data-driven methods

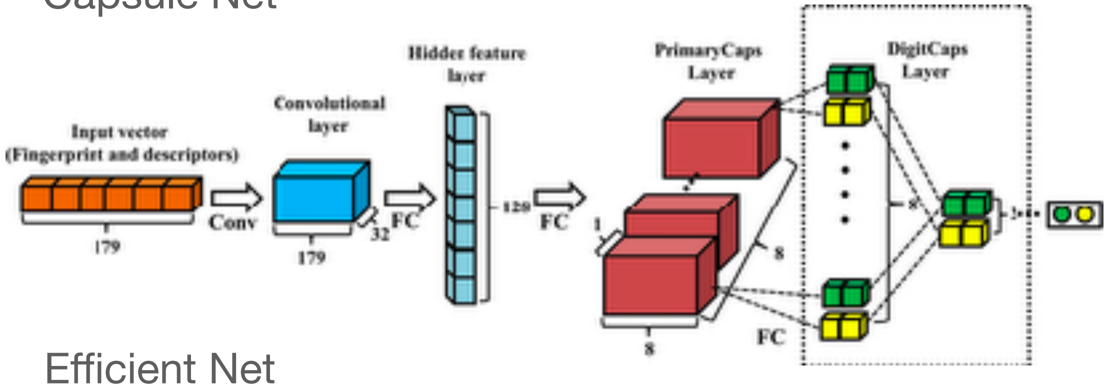


Deep neural network

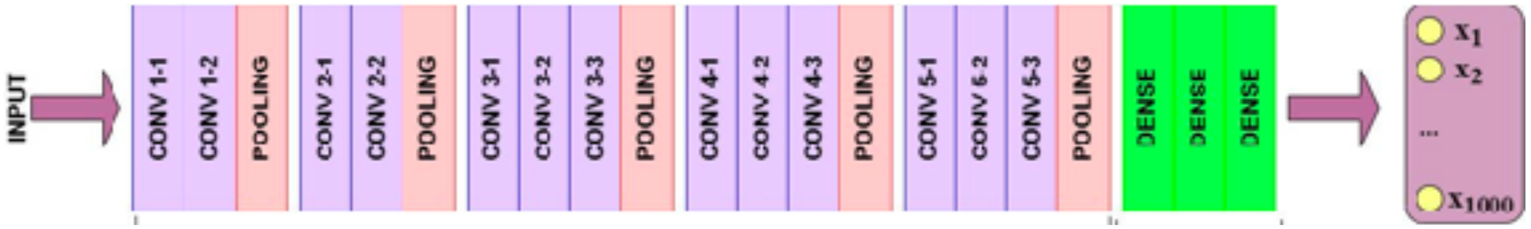
Resent



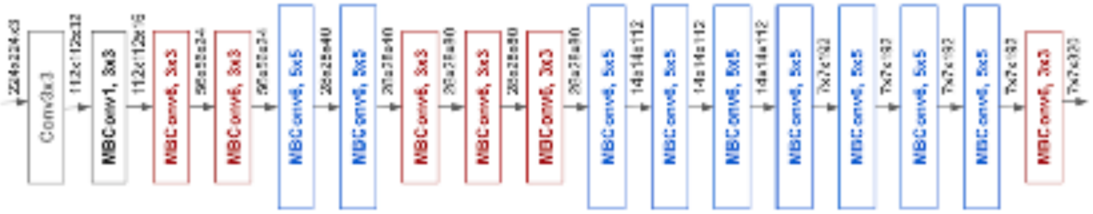
Capsule Net



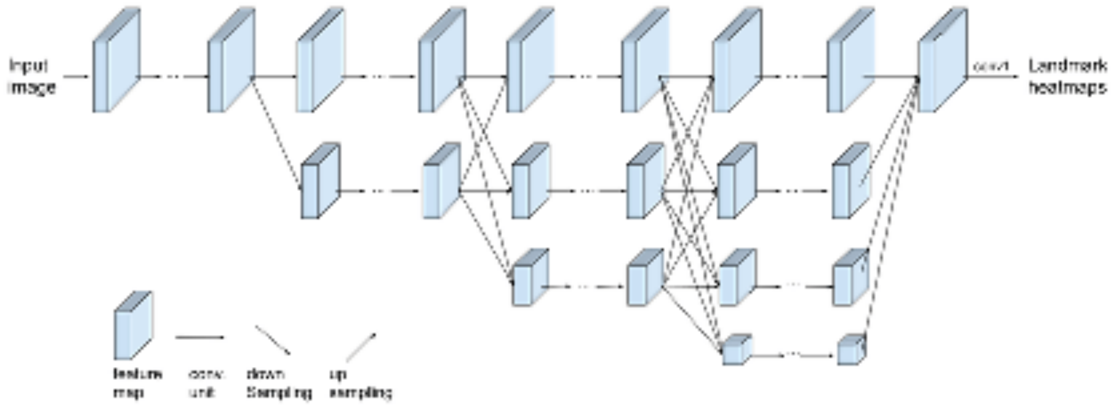
VGG



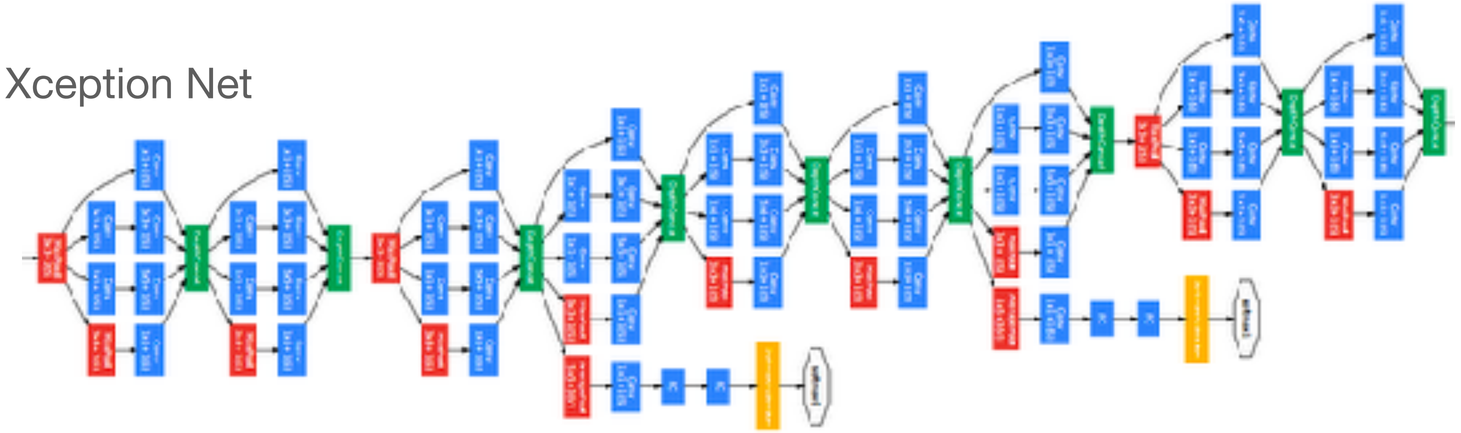
Efficient Net



HRNet



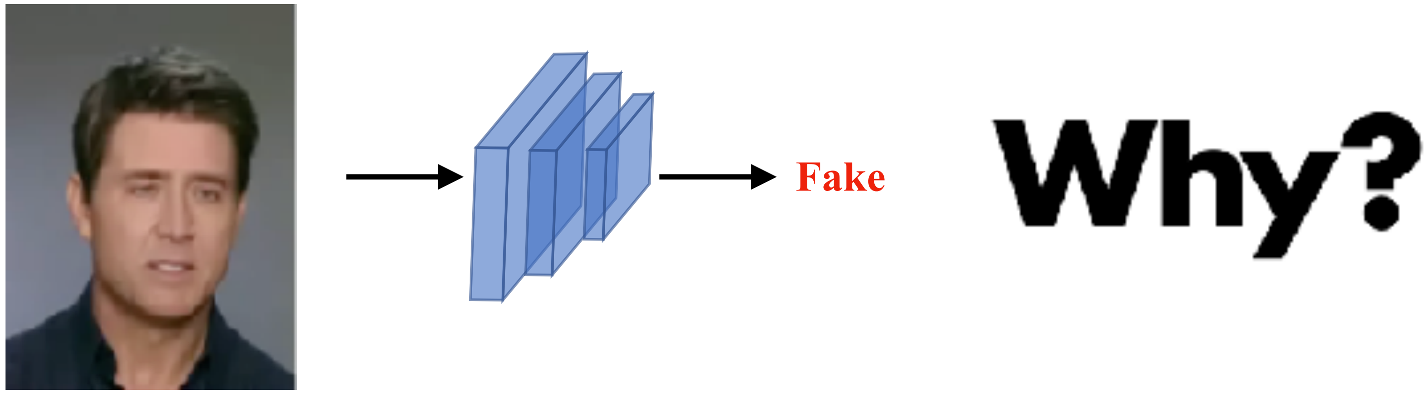
Xception Net



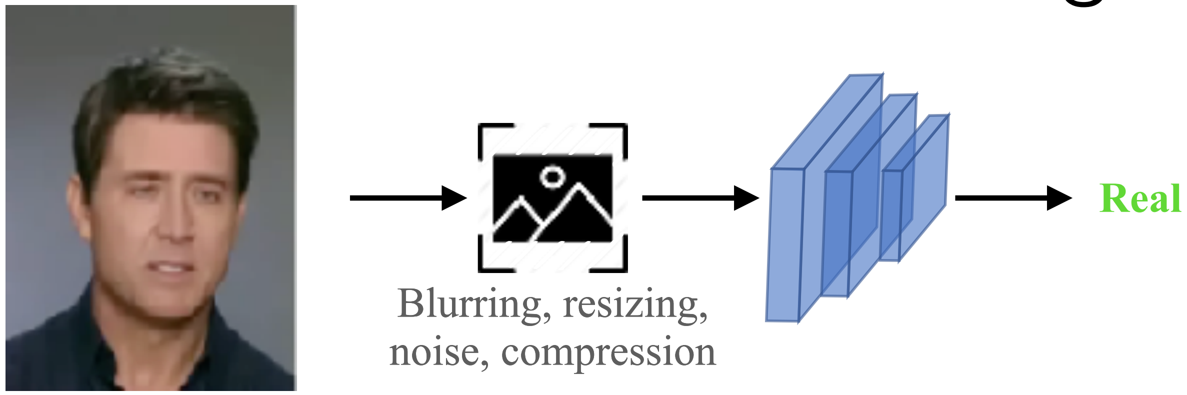
© Siwei Lyu, All rights reserved.

Data-driven methods: challenges

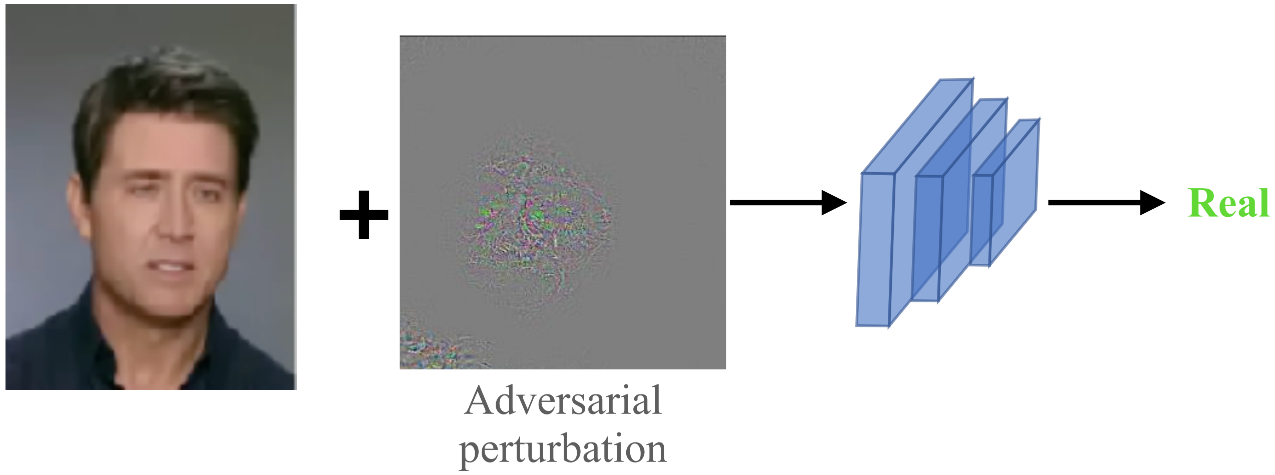
- Lack of explainability



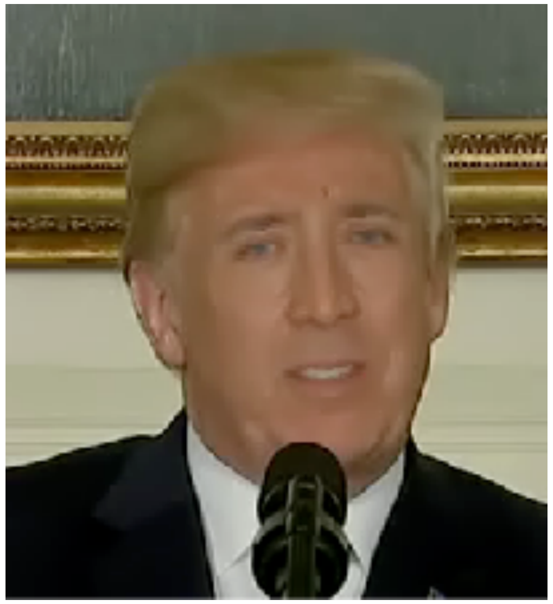
- Lack of robustness to laundering operations



- Susceptibility to anti-forensics

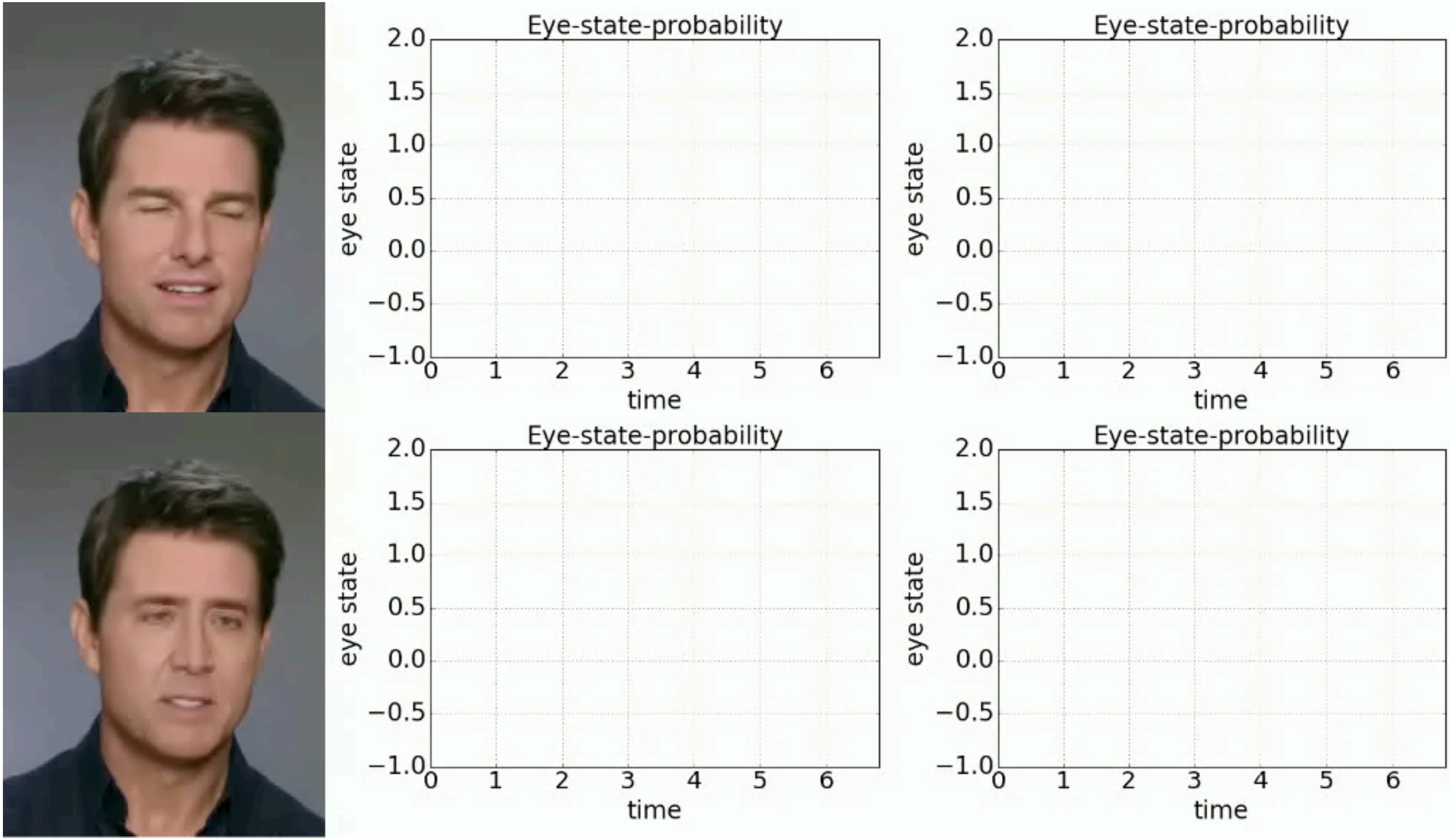


Physiological cues — blinking



They do not blink!

Physiological cues — blinking



Yuezun Li, Ming-Ching Chang and Siwei Lyu. **In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking**. In WIFS 2018

Physiology: Pupil shapes

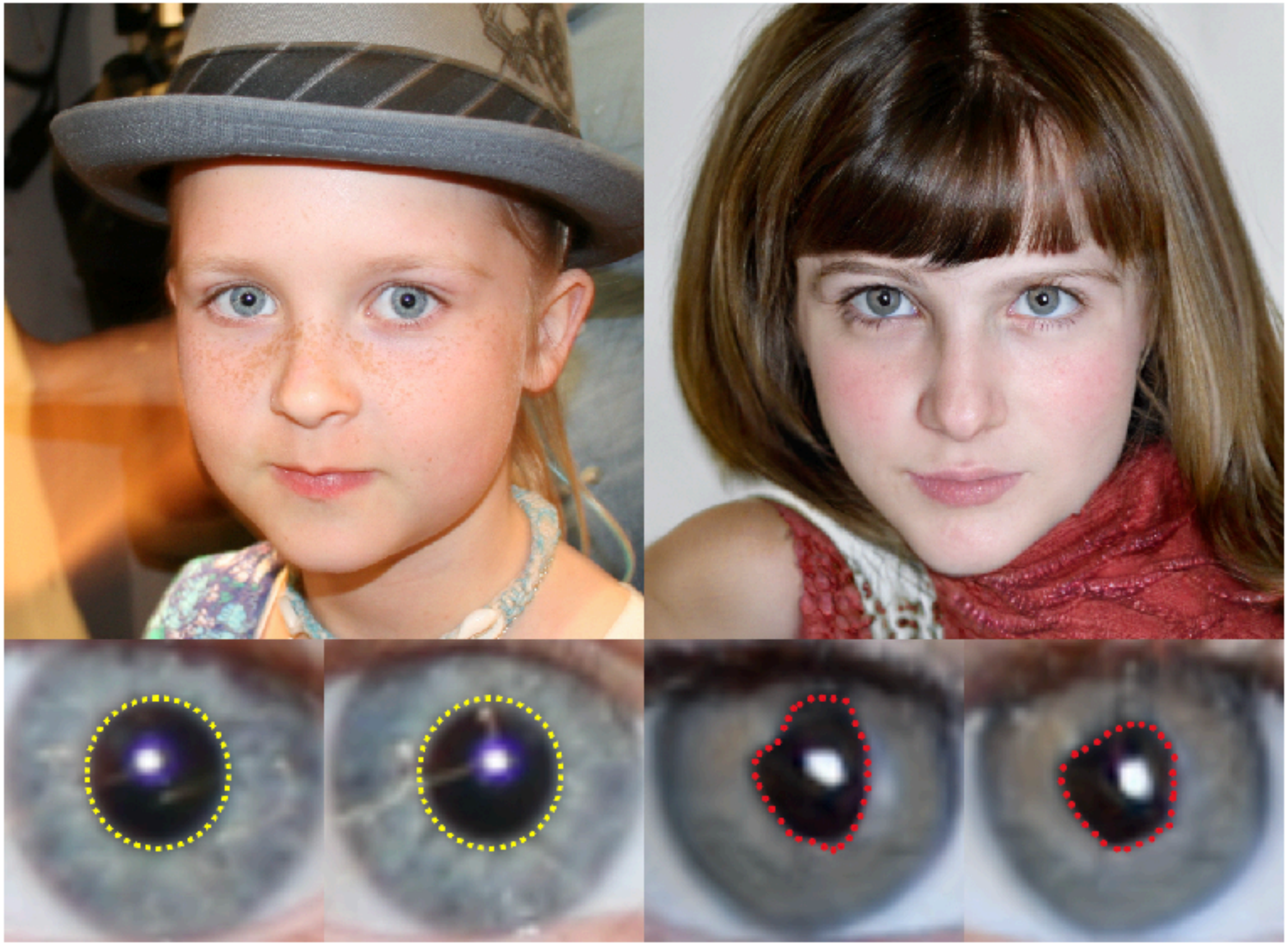
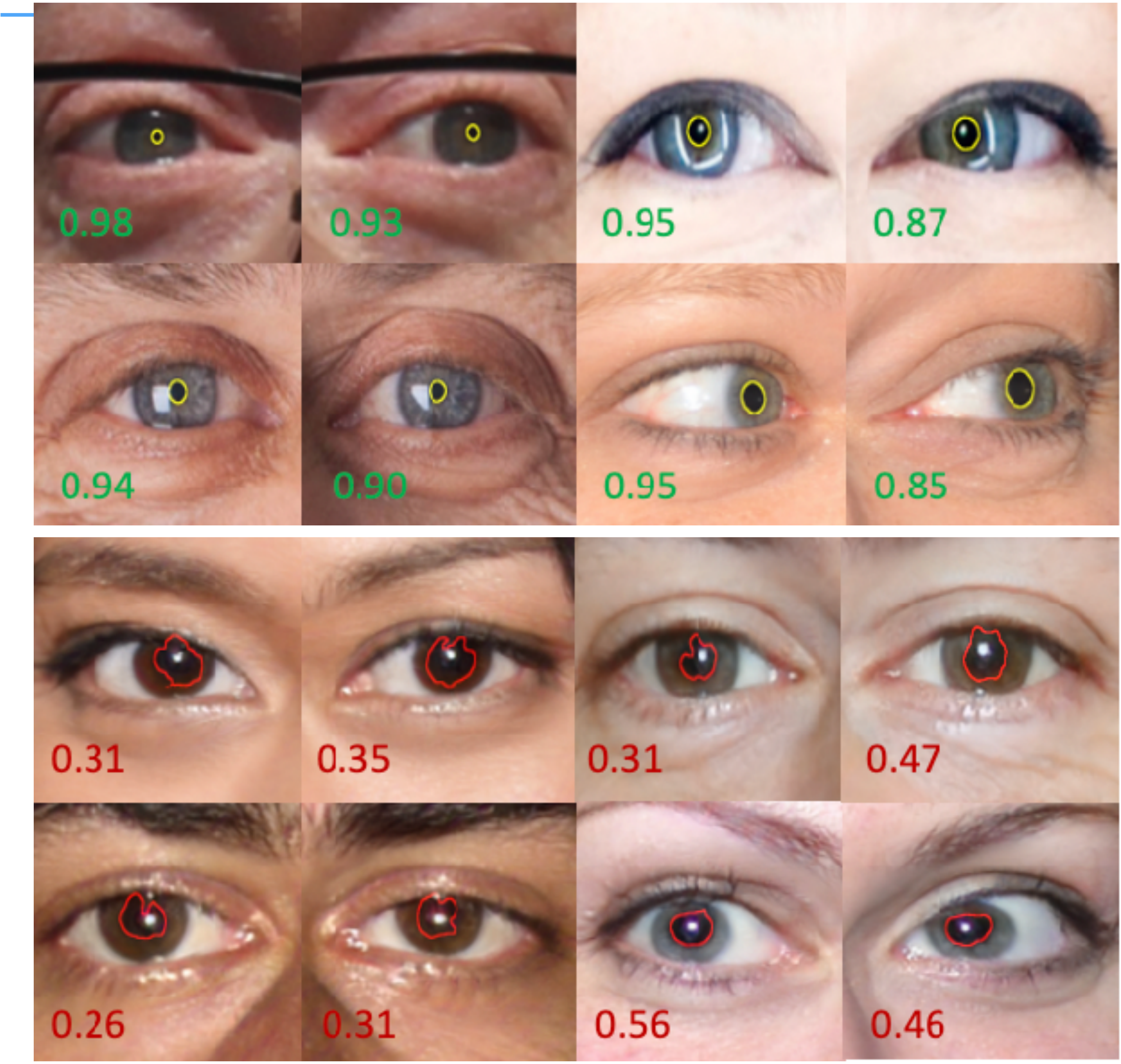
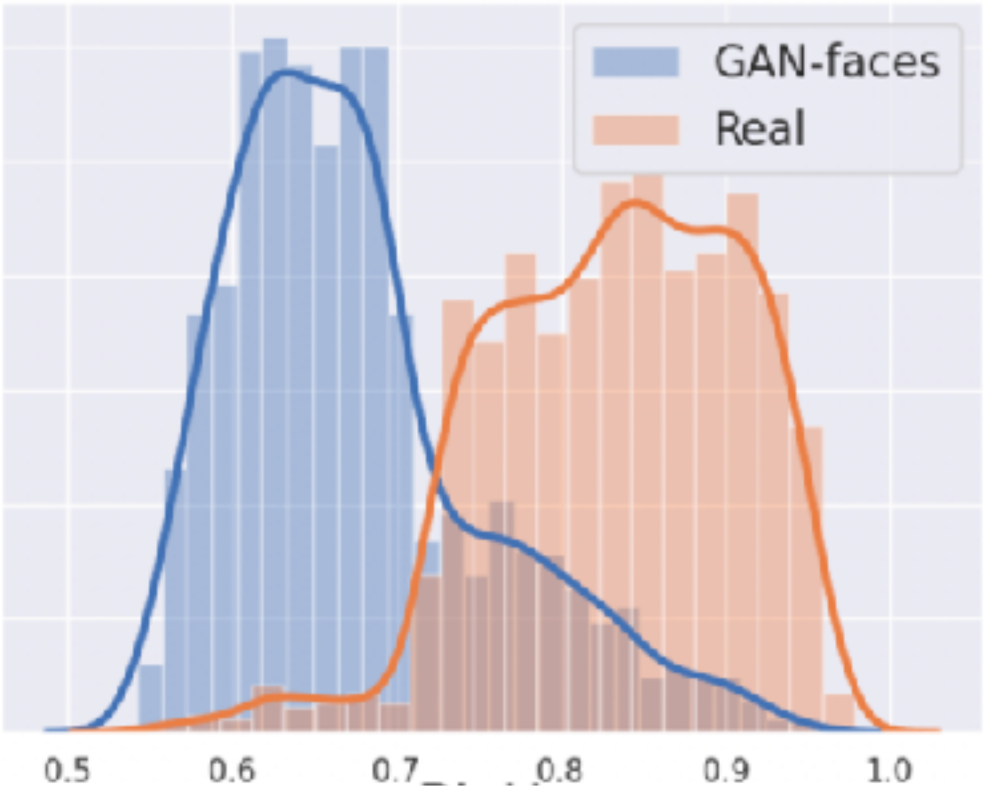
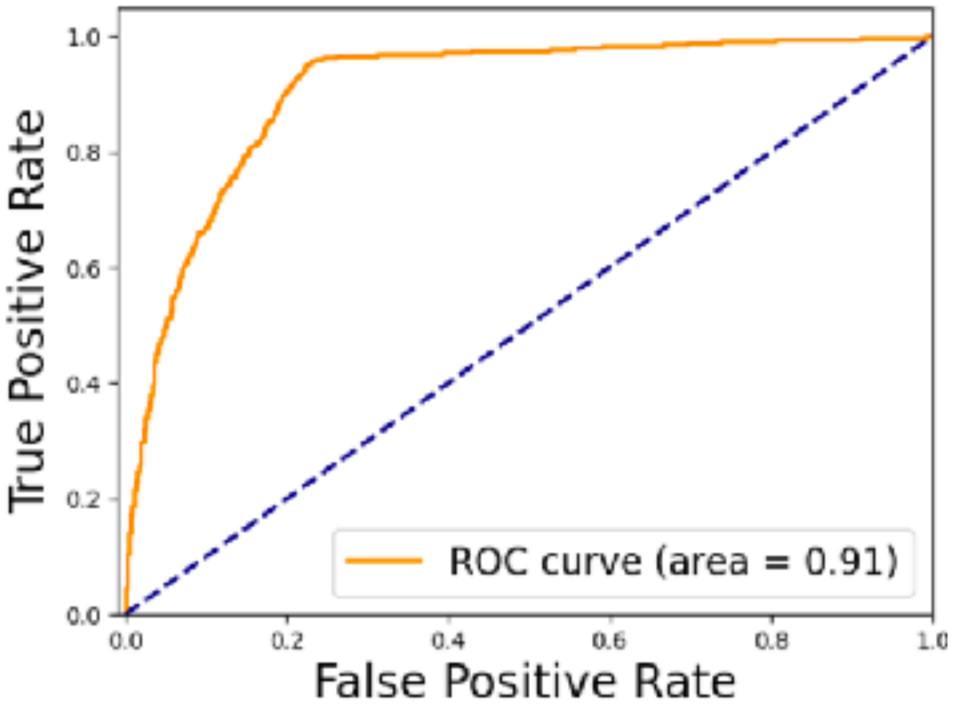


Image source: www.thispersondoesnotexist.com

Physiology: Pupil shapes



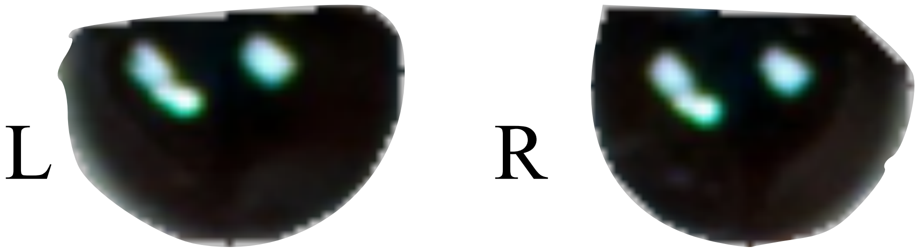




Hui Guo, Shu Hu, Xin Wang, Ming-Ching Chang, and Siwei Lyu. **Eyes Tell All: Irregular Pupil Shapes Reveal Gan-Generated Faces**. In ArXiv 2021

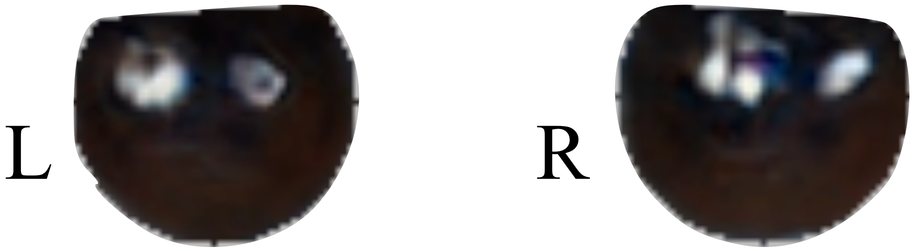
Physical cues — corneal reflections

Real

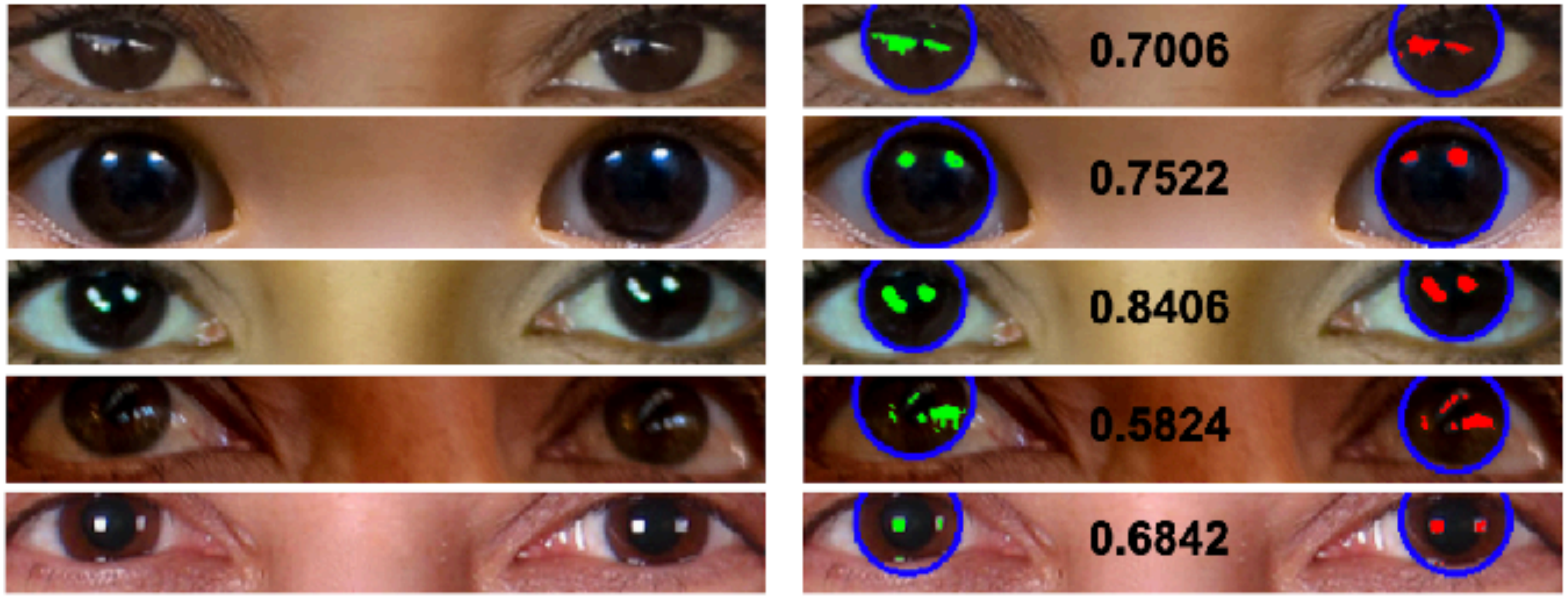


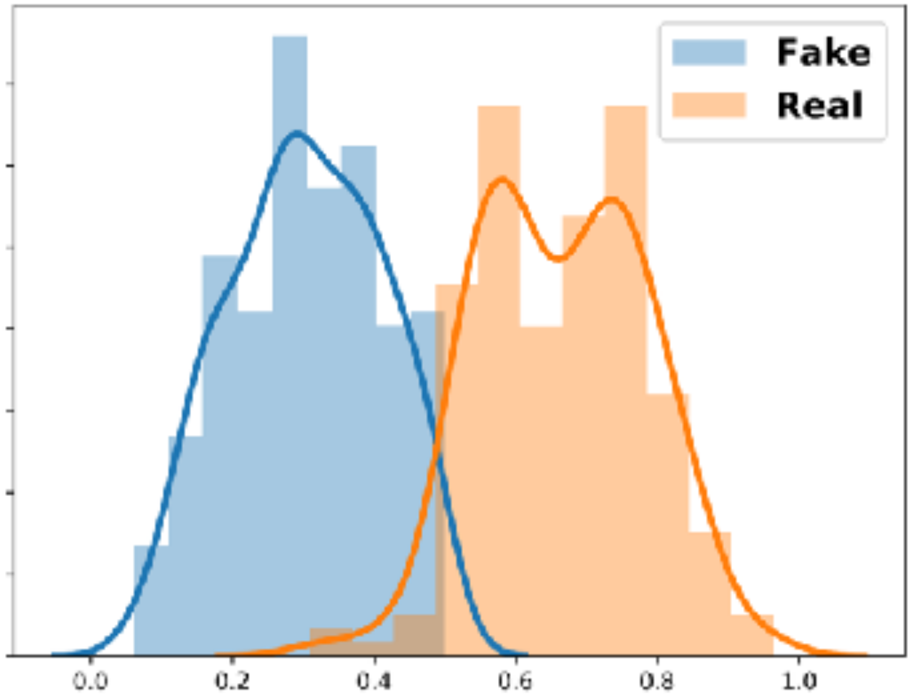
StyleGAN2

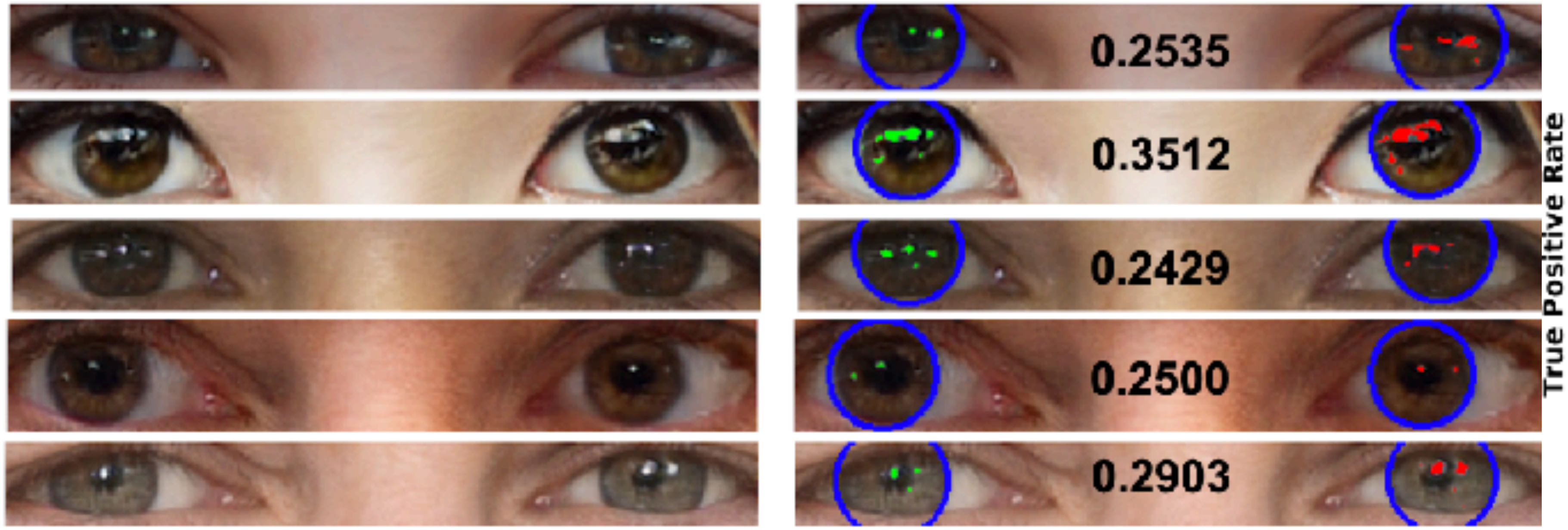
Image source: www.thispersondoesnotexist.com

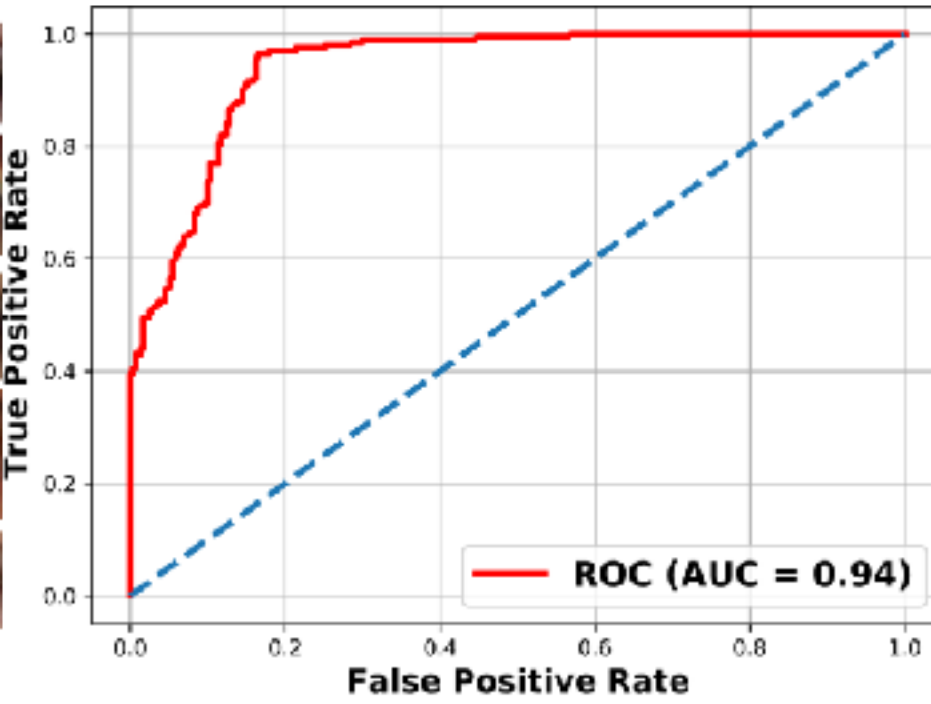


Results



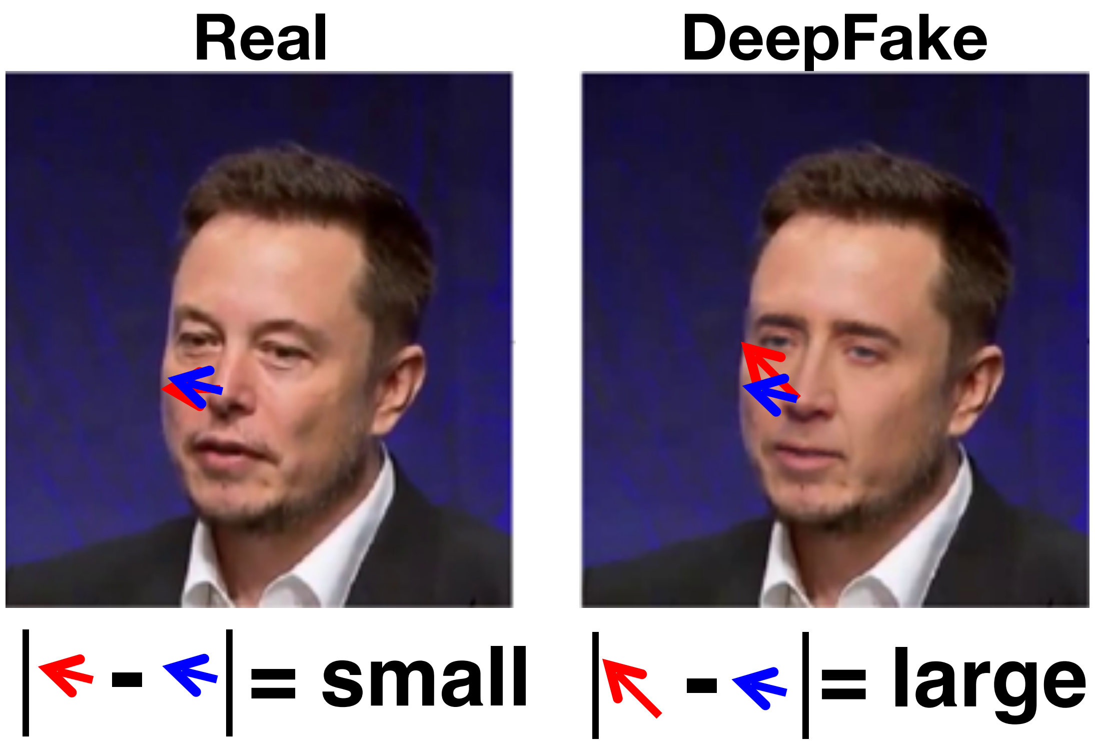


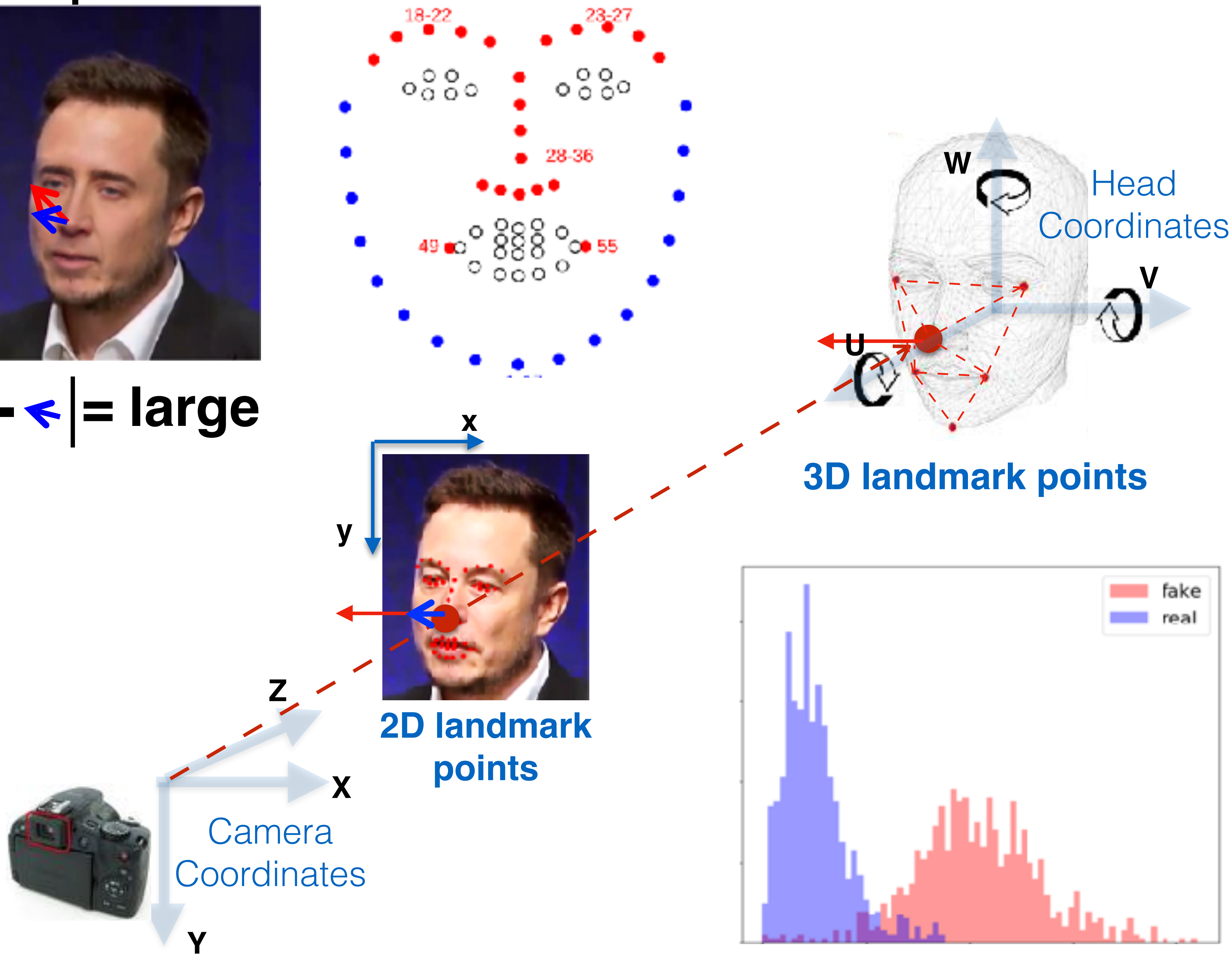




Shu Hu, Yuezun Li, and Siwei Lyu. **Exposing GAN-synthesized Faces using Inconsistent Corneal Specular Reflection Highlights**. In ICASSP 2021

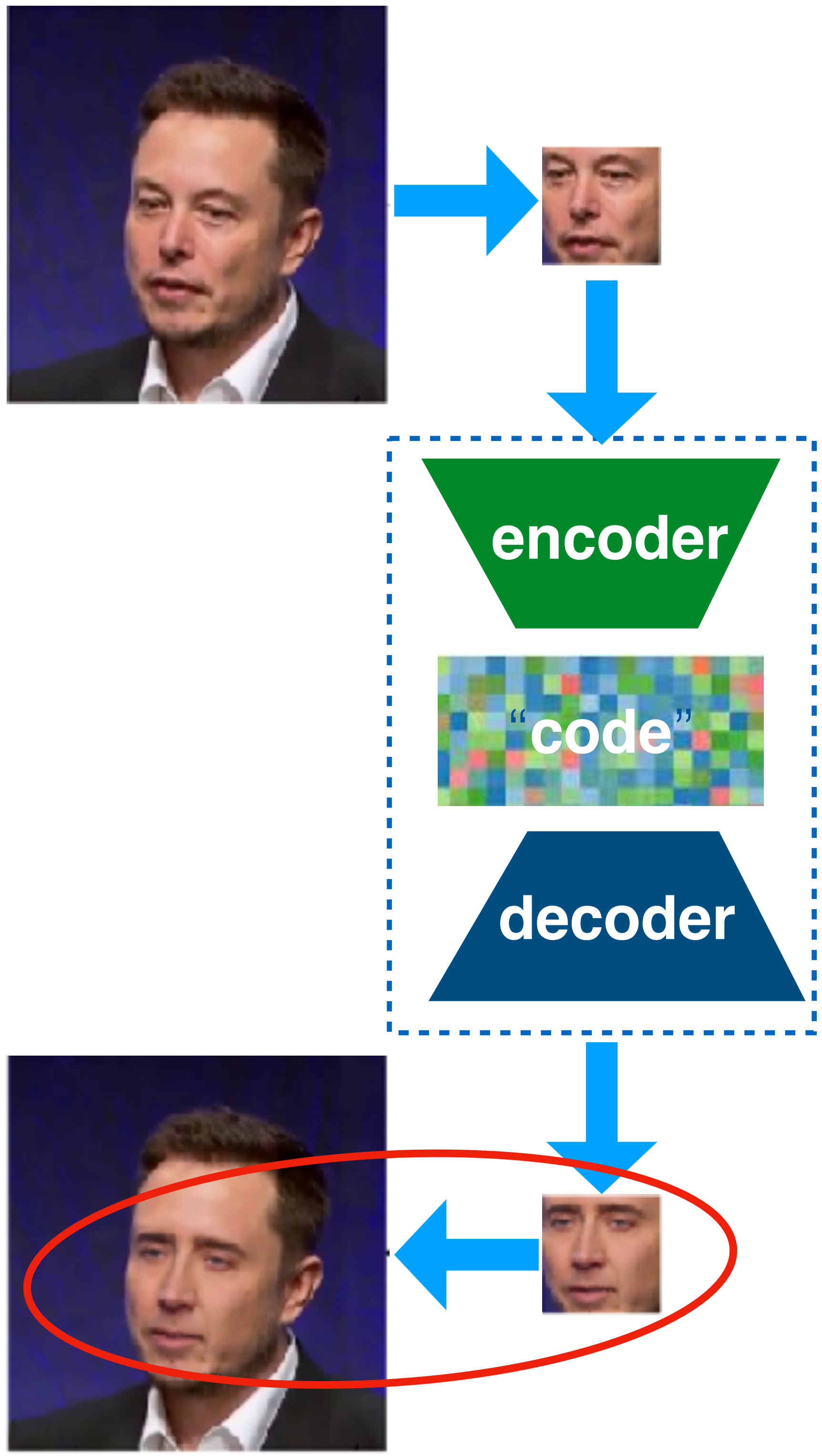
Physical cues — face orientations

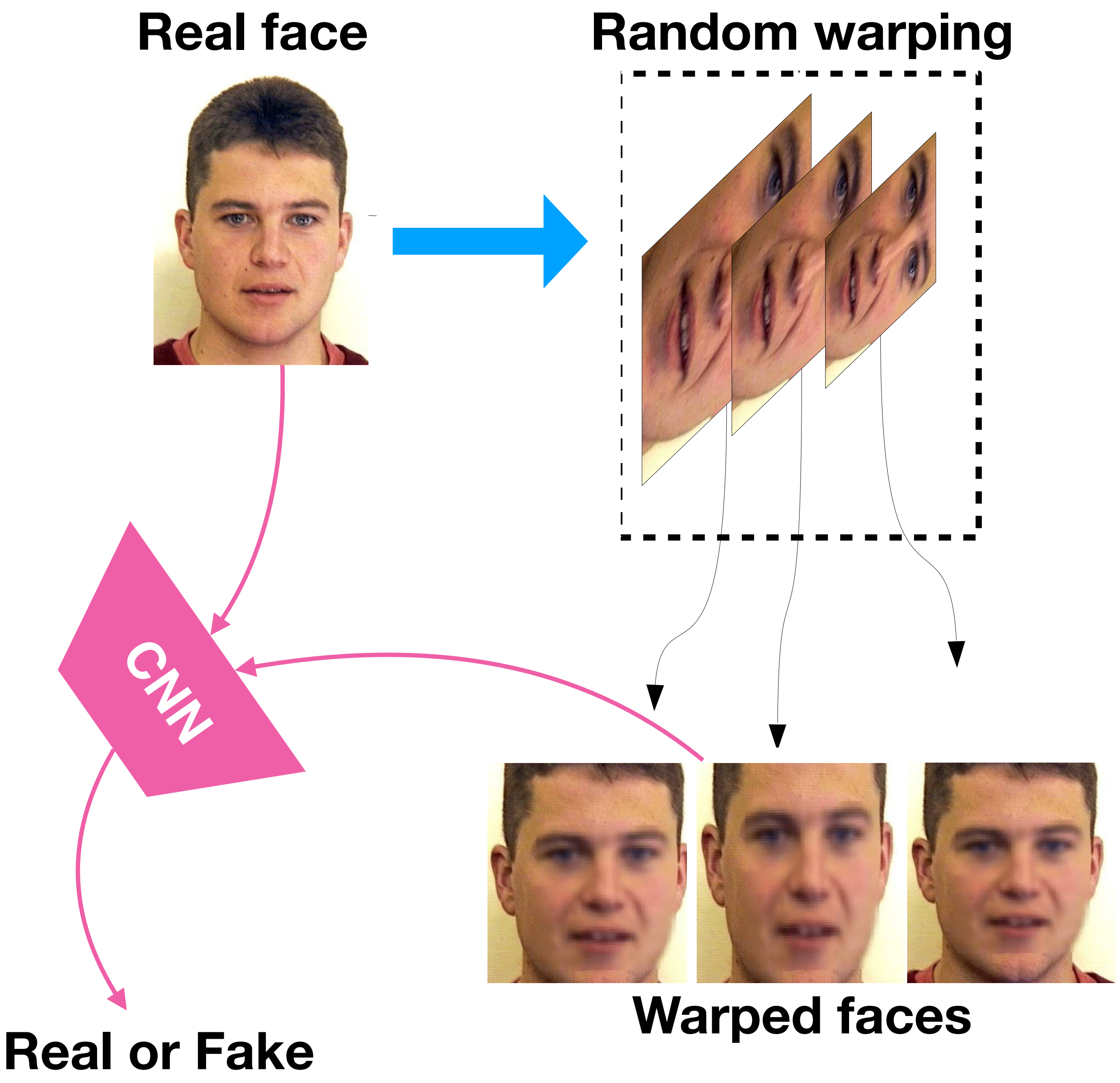




Xin Yang, Yuezun Li and Siwei Lyu. Exposing Deep Fakes Using Inconsistent Head Poses. ICASSP 2019

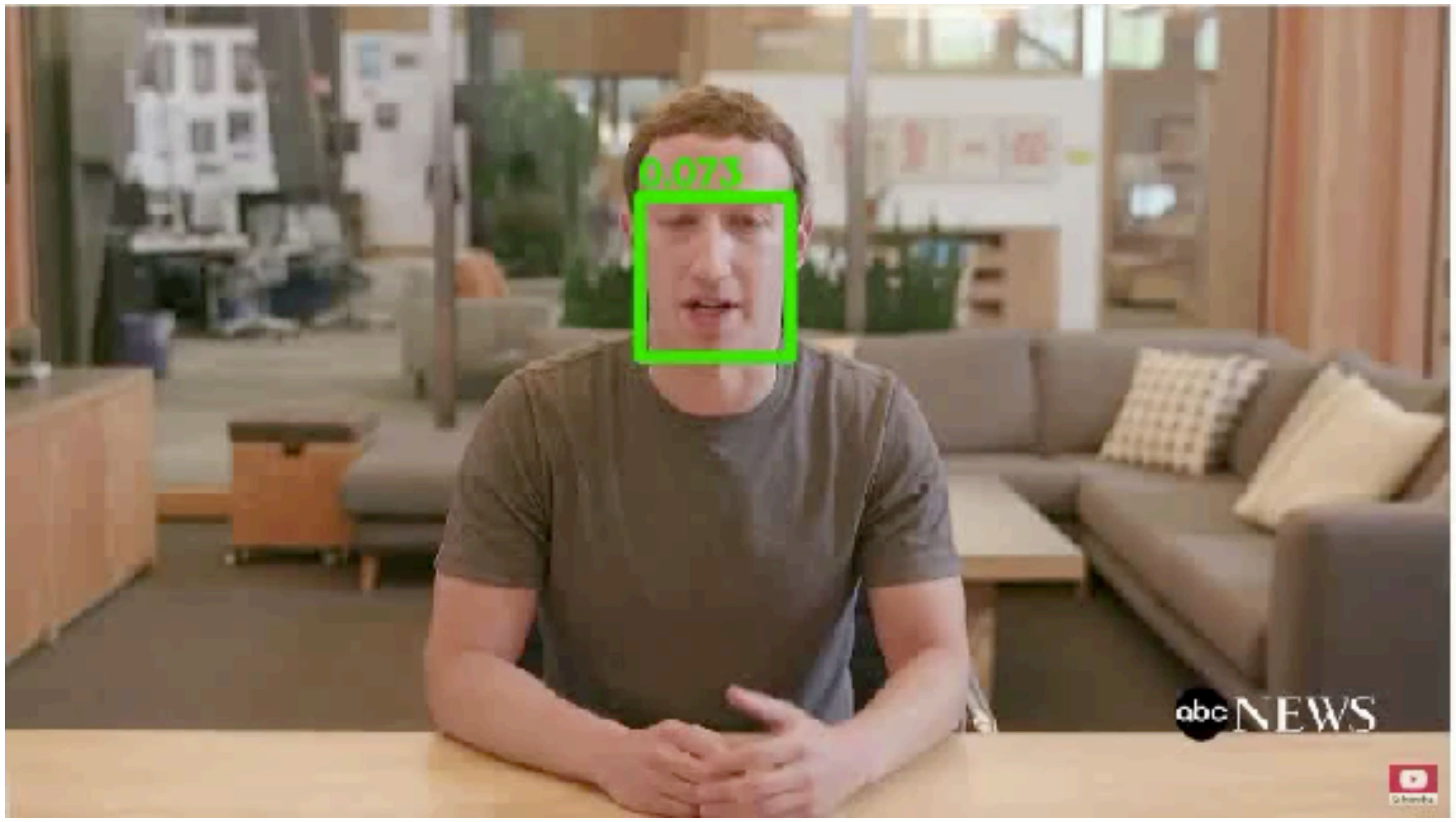
Signal cues — splicing traces



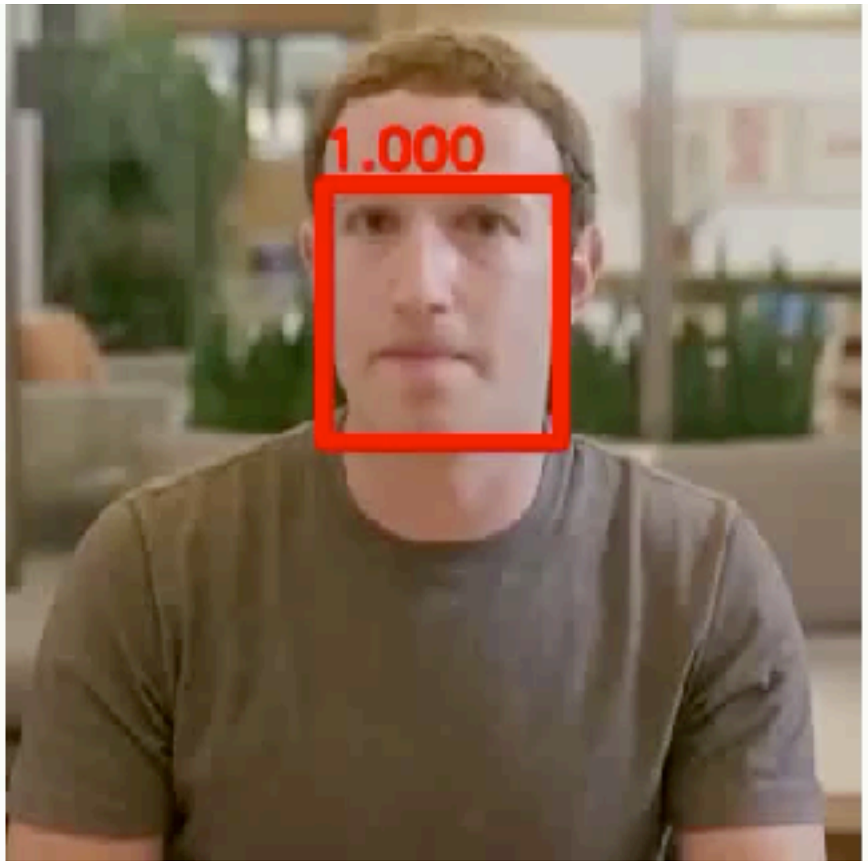


Yuezun Li and Siwei Lyu. Exposing Deep Fake Videos By Detecting Face Warping Artifacts. In CVPRW 2019

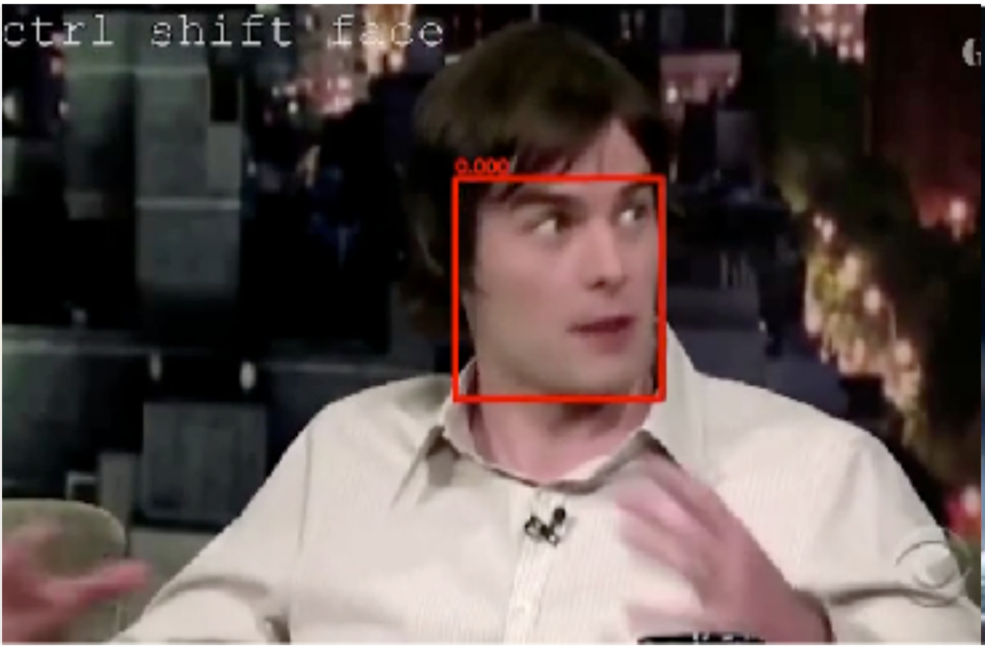
Exposing DeepFakes: results



real



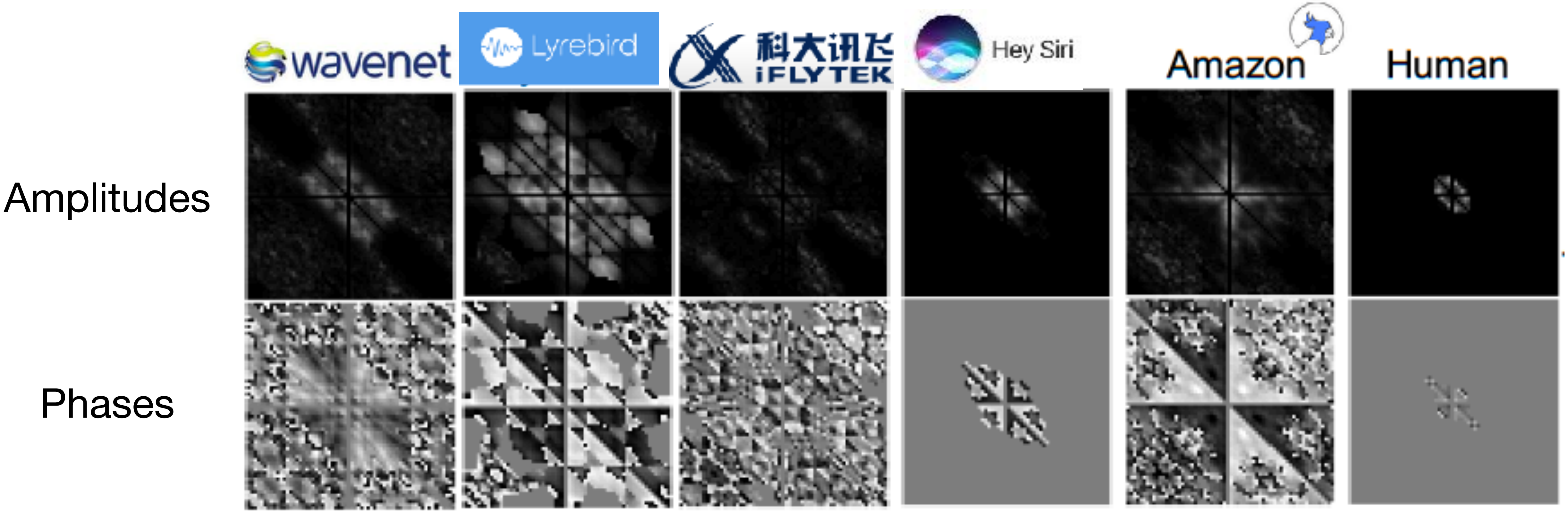
Canny AI



Ctrl Shift Face

ZAO

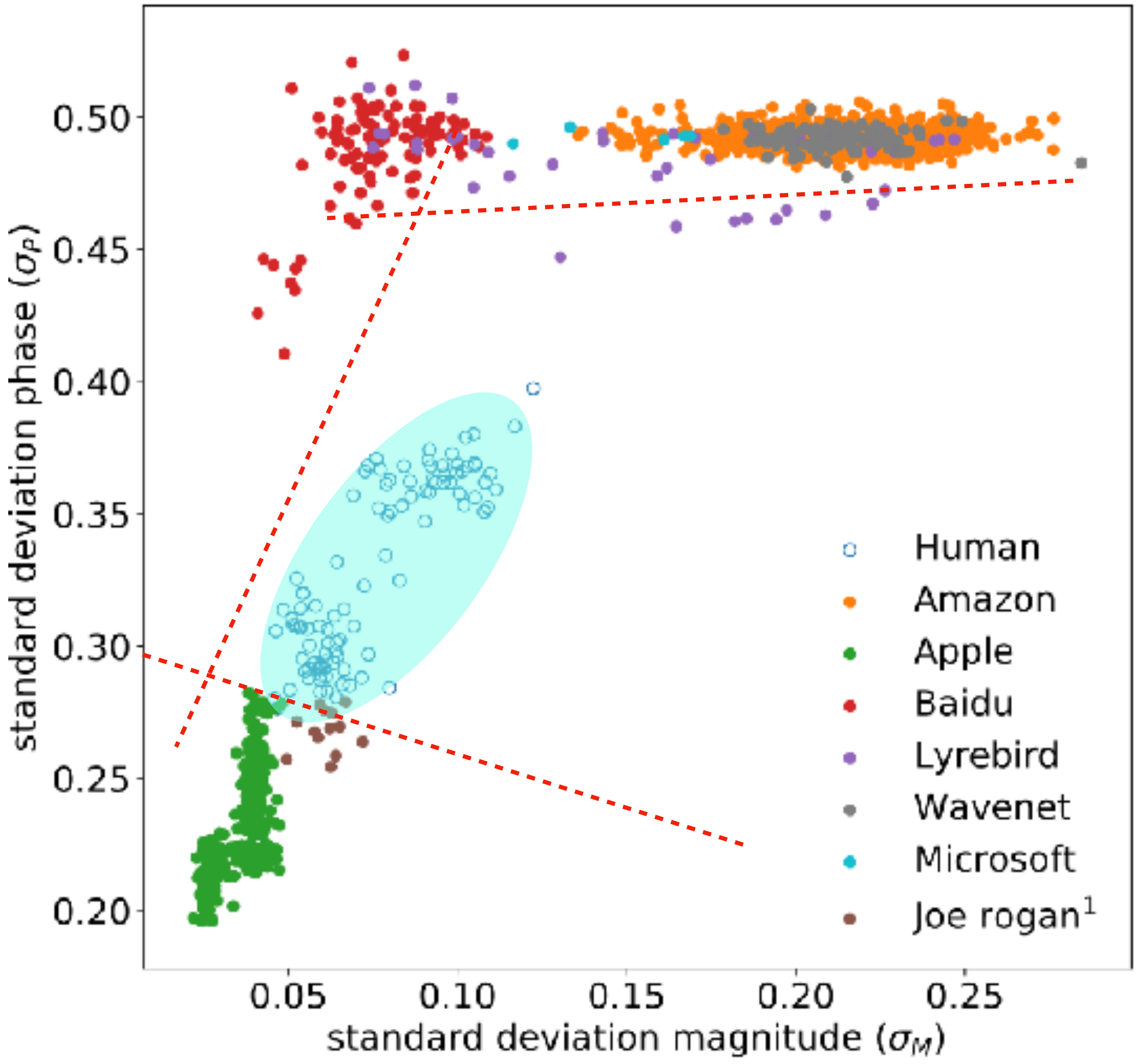
Bi-spectra of human and AI-synth voices

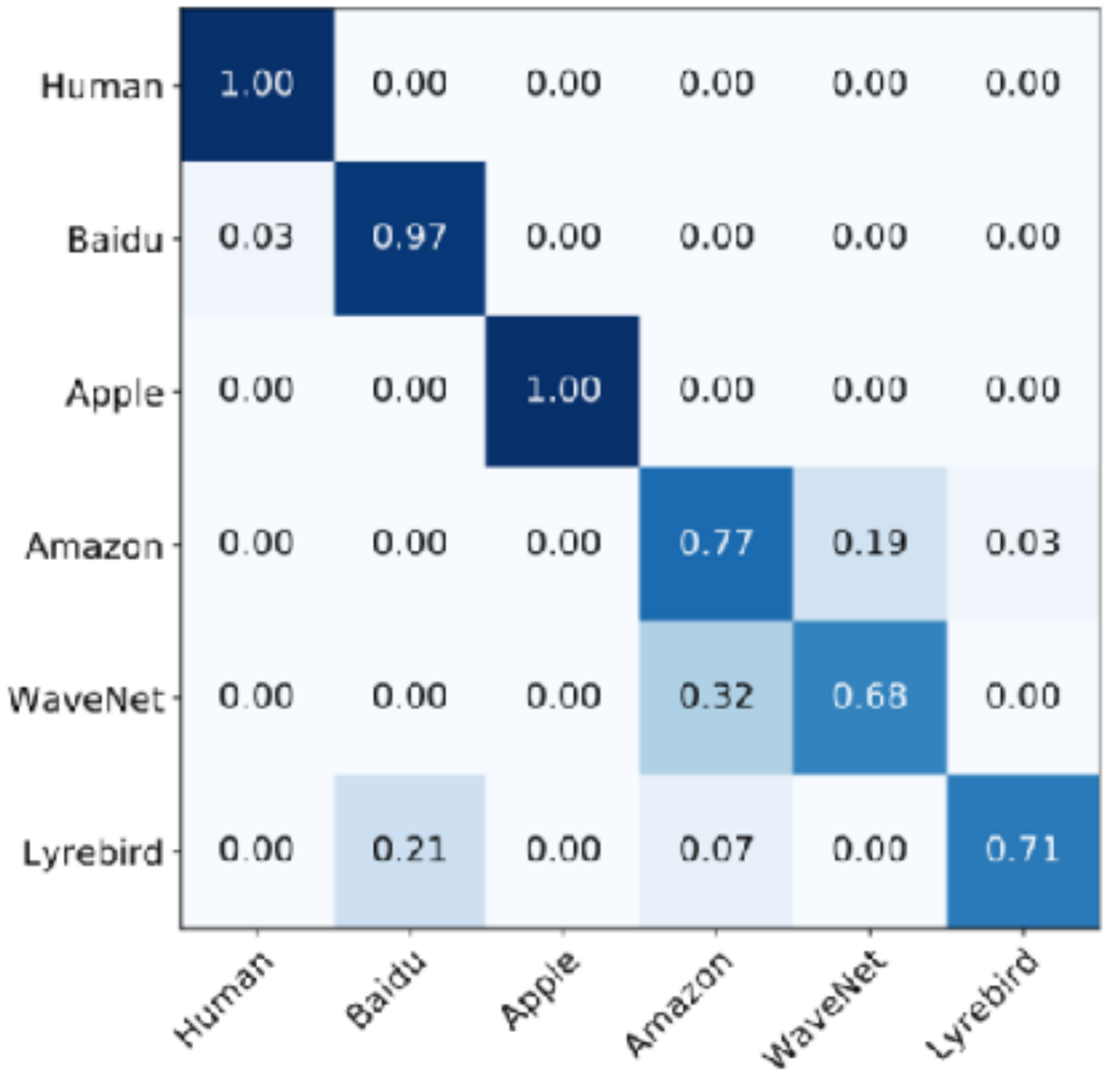


Ehab ElBadawy, Siwei Lyu, and Hany Farid. Detecting AI-Synthesized Speech Using Bispectral Analysis, in CVPRW MediFor 2019

Classification results

Classification results based on multi-class logistic regression

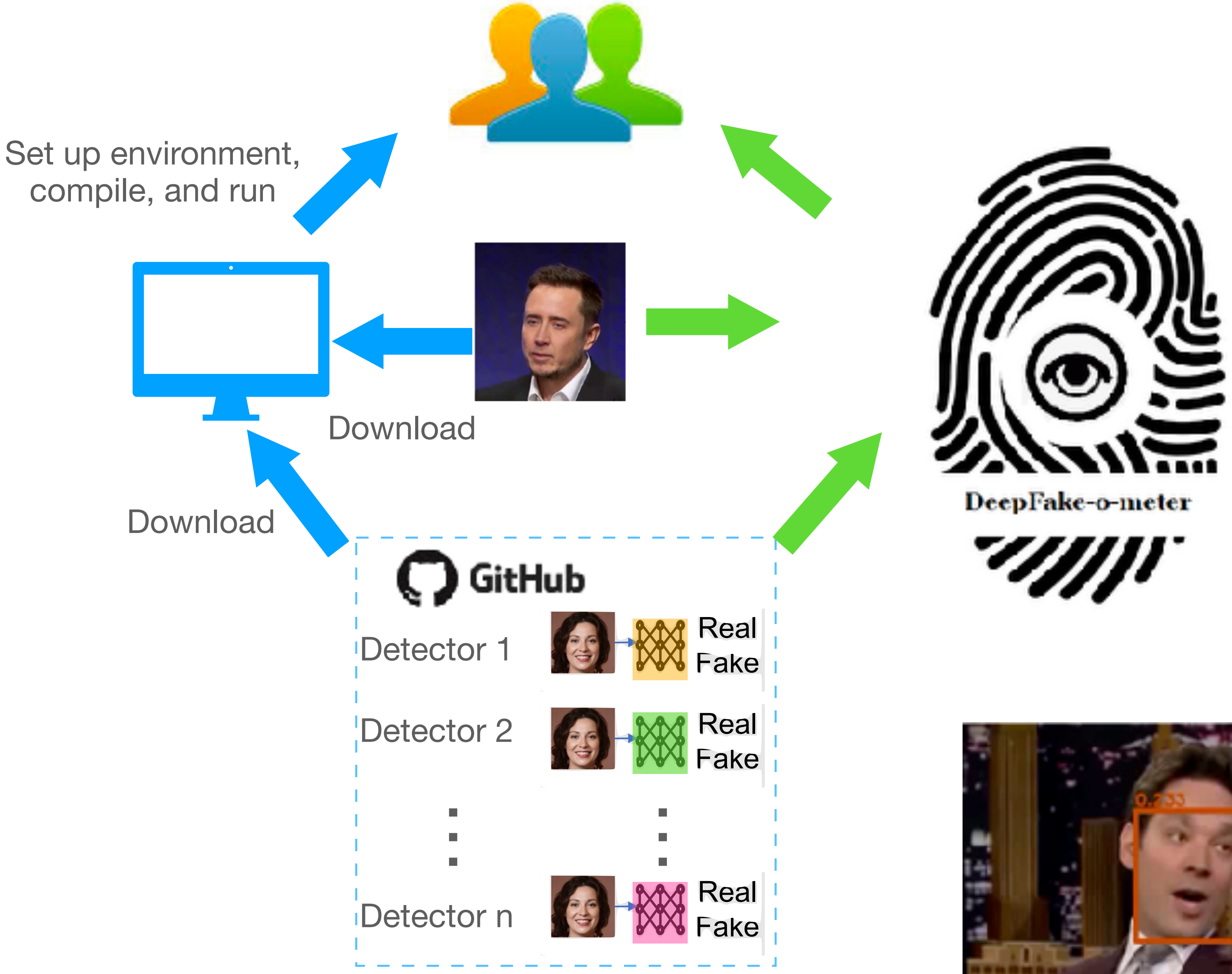




	Human	Baidu	Apple	Amazon	WaveNet	Lyrebird
Human	1.00	0.00	0.00	0.00	0.00	0.00
Baidu	0.03	0.97	0.00	0.00	0.00	0.00
Apple	0.00	0.00	1.00	0.00	0.00	0.00
Amazon	0.00	0.00	0.00	0.77	0.19	0.03
WaveNet	0.00	0.00	0.00	0.32	0.68	0.00
Lyrebird	0.00	0.21	0.00	0.07	0.00	0.71

Ehab ElBadawy, Siwei Lyu, and Hany Farid. Detecting AI-Synthesized Speech Using Bispectral Analysis, in CVPRW MediFor 2019

DeepFake-o-meter

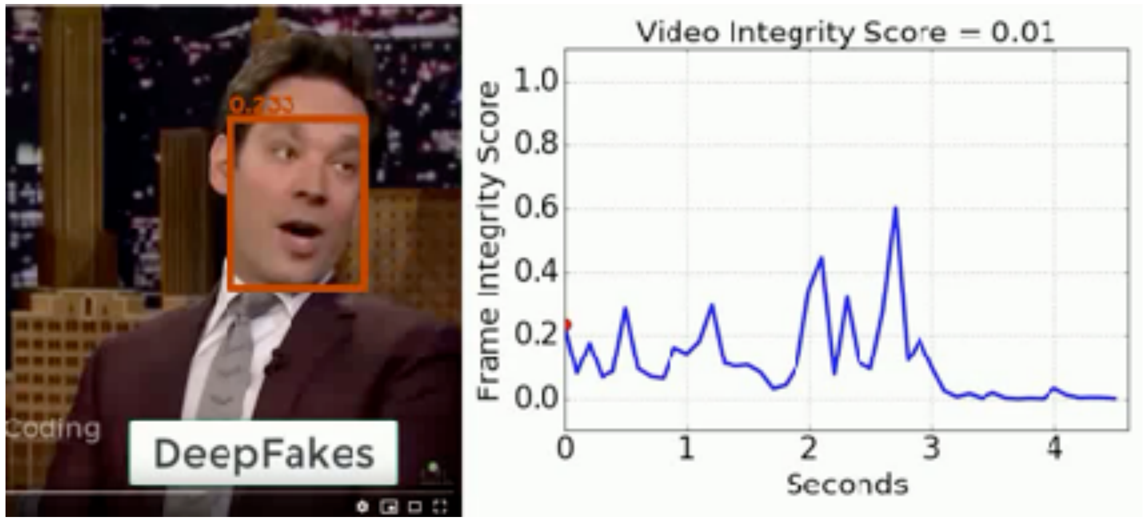


Reality Defender

FIXFAKE

FakeNetAI

sensity

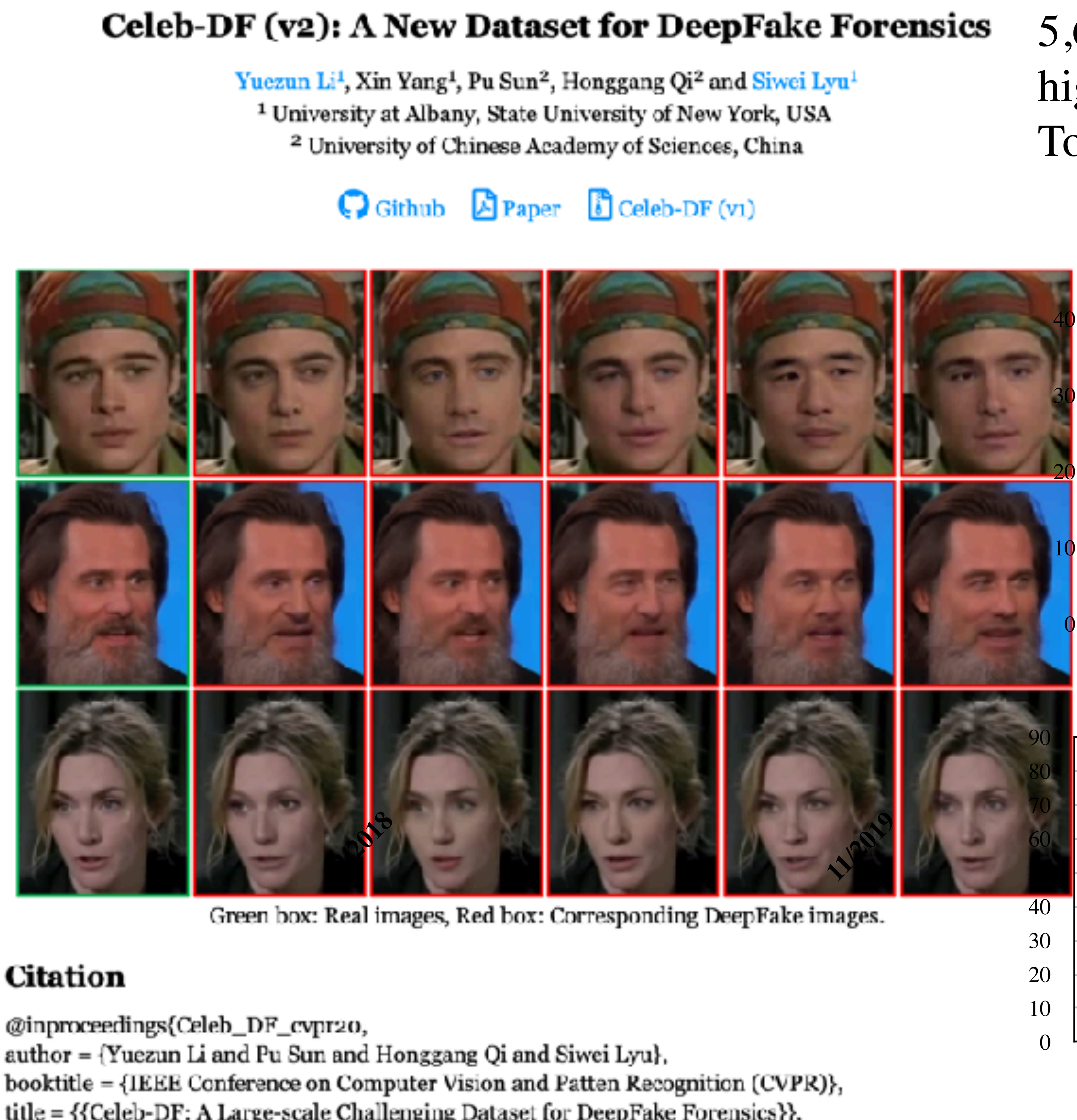


http://zinc.cse.buffalo.edu/ubmdfl/deep-o-meter/

Yuezun Li, Cong Zhang, Pu Sun, Lipeng Ke, Yan Ju, Honggang Qi and Siwei Lyu. **DeepFake-o-meter: An Open Platform for DeepFake Detection**. In International Conference on Systematic Approaches to Digital Forensic Engineering (SADFE), 2021

The Celeb-DF dataset

http://www.cse.buffalo.edu/~siweilyu/celeb-deepfakeforensics.html



Celeb-DF (v2): A New Dataset for DeepFake Forensics

Yuezun Li[1], Xin Yang[1], Pu Sun[2], Honggang Qi[2] and Siwei Lyu[1]
[1] University at Albany, State University of New York, USA
[2] University of Chinese Academy of Sciences, China

Github Paper Celeb-DF (v1)

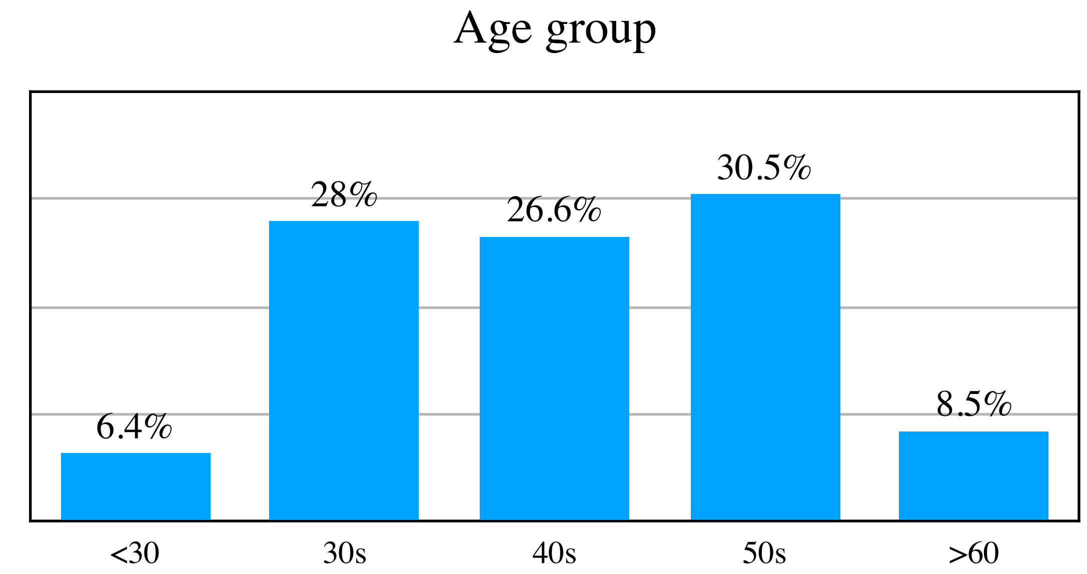
Green box: Real images, Red box: Corresponding DeepFake images.

Citation

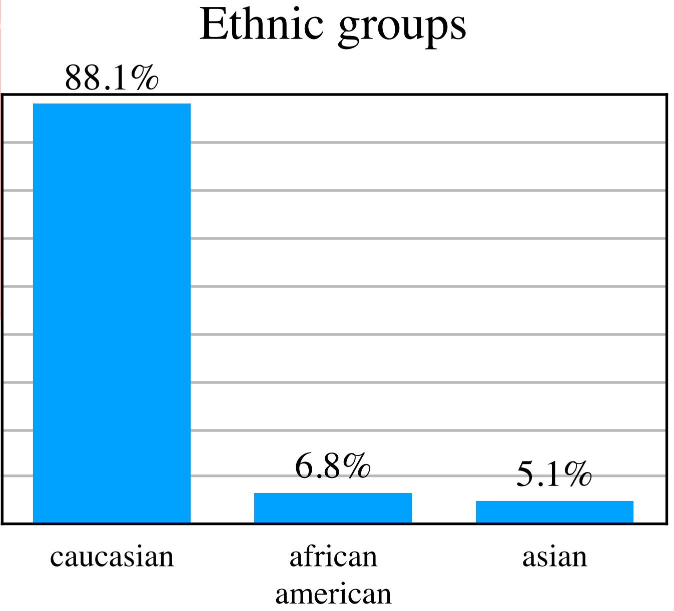
```
@inproceedings{Celeb_DF_cvpr20,
author = {Yuezun Li and Pu Sun and Honggang Qi and Siwei Lyu},
booktitle = {IEEE Conference on Computer Vision and Patten Recognition (CVPR)},
title = {{Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics}},
```

5,600 video clips with 2 million frames of high-quality synthetic DeepFake videos. Total downloads > 2,000 since Nov. 2019

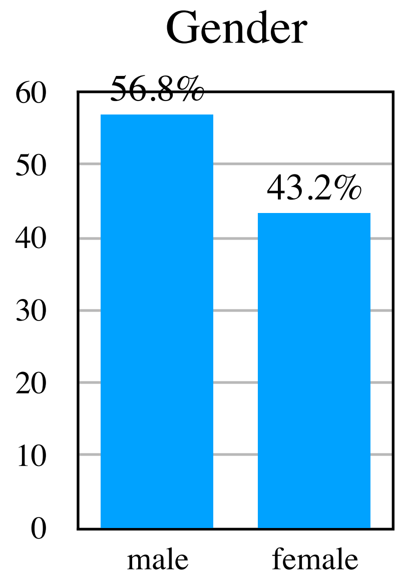
Age group



Ethnic groups

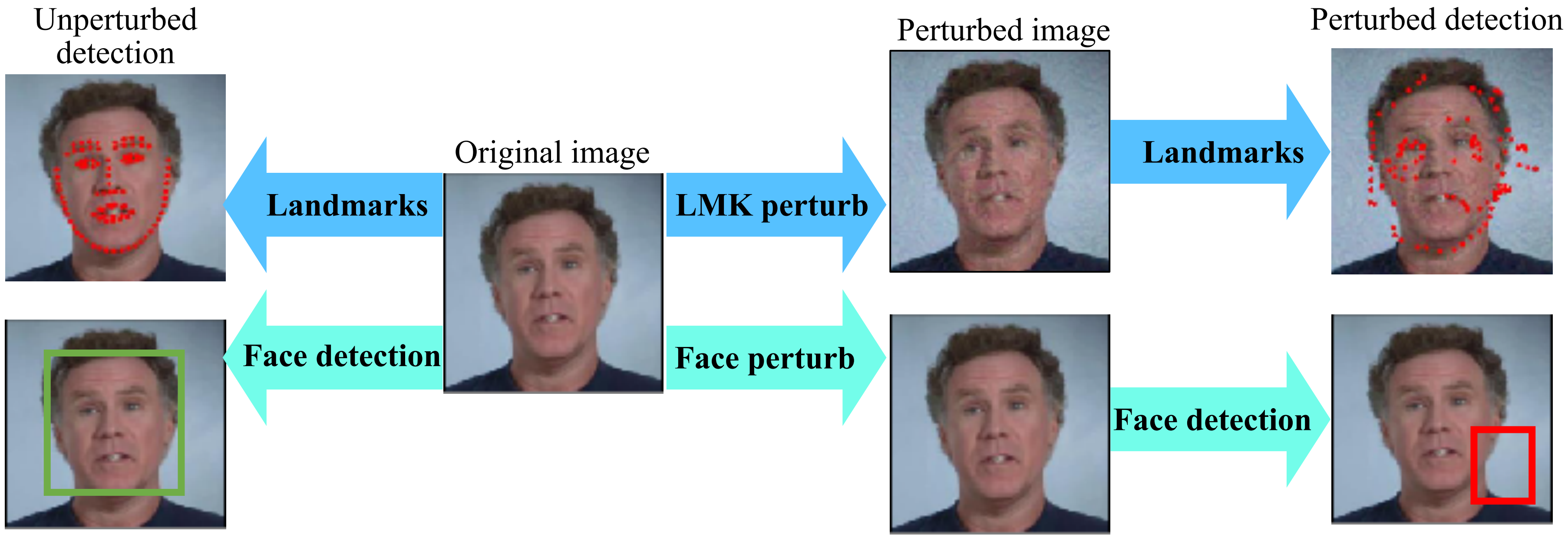


Gender



Beyond detection

Pollute the training data to disrupt model training



Unperturbed detection

Landmarks

Original image

LMK perturb

Perturbed image

Landmarks

Perturbed detection

Face detection

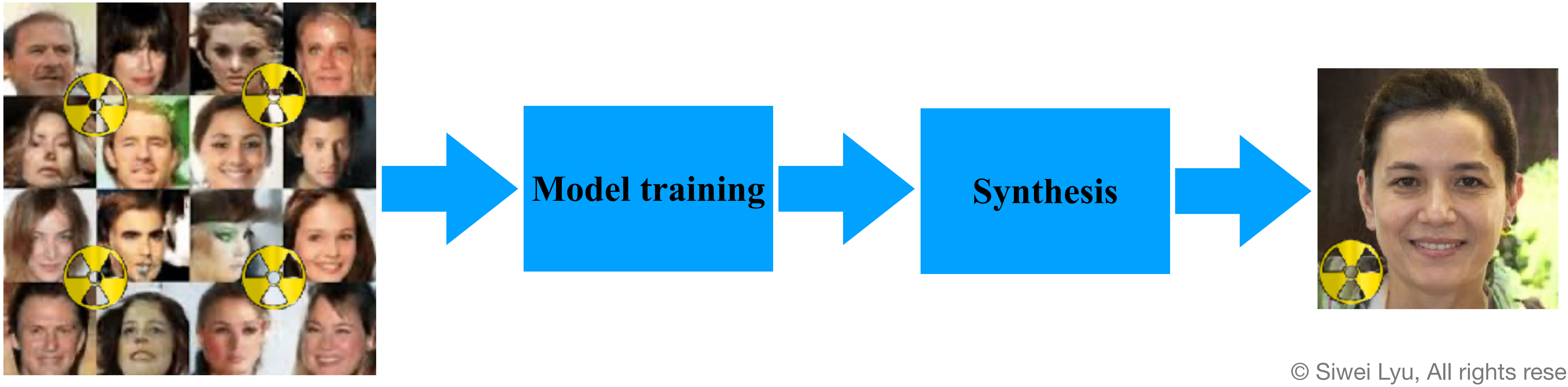
Face perturb

Face detection

Yuezun Li, Xin Yang, Baoyuan Wu and Siwei Lyu. **Hiding Faces In Plain Sight: Disrupting AI Face Synthesis with Adversarial Perturbations**. In ArXiv 2019

Pu Sun, Yuezun Li, Honggang Qi, and Siwei Lyu. **Landmark Breaker: Obstructing DeepFake By Disturbing Landmark Extraction,** in WIFS 2020

Make fake media traceable [Yu et.al. 2021]



Model training

Synthesis

The future of DeepFakes: Artificial Reality

Credit: work from UC Berkeley, U. Washington, MPI, and Nvidia Research

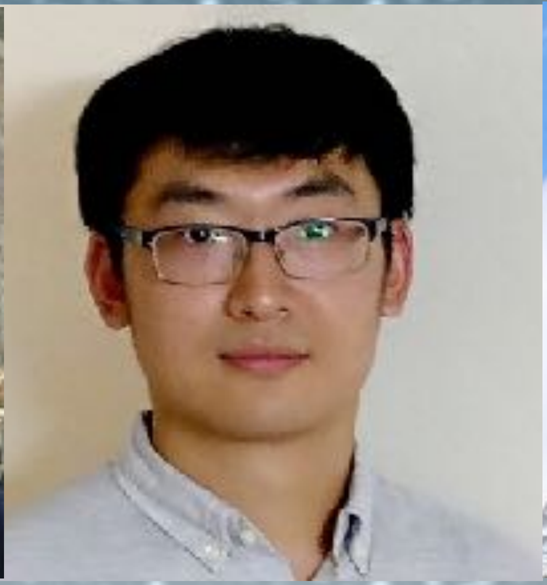
Summary

- AI algorithms facilitate creating AI synthesized media that can cause real damage.
- Forensic techniques aim to detect and obstruct AI synthesized fake media.
- Do not forget the “ShallowFakes”!
- The competition between forgery making and forensics is a perpetual cat-and-mouse game.

- Fighting fake media and disinformation is a community effort, involving academia, social platforms, government agencies, media and online users — all hands on deck!

Thank You!

Prof. Hany Farid
UC Berkeley



Yuezun Li
Post-doc

Shu Hu
Ph.D. student

Xin Yang
M.S. student

Ehab Albadaway
Ph.D. student

Work presented in this talk is sponsored by the following agencies.

Opinions and comments in this talk are not associated with any sponsoring agency.