What are deepfakes, and how can we detect them?

Shannon Gallagher and Dominic Ross

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution

Copyright 2022 Carnegie Mellon University.

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

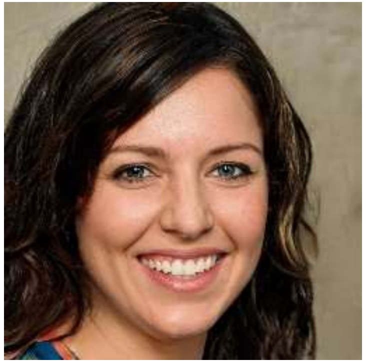
[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

This material may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other use. Requests for permission should be directed to the Software Engineering Institute at permission@sei.cmu.edu.

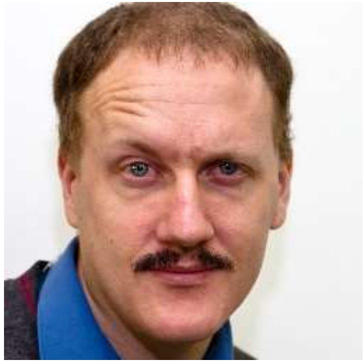
Carnegie Mellon® and CERT® are registered in the U.S. Patent and Trademark Office by Carnegie Mellon University.

DM22-0713

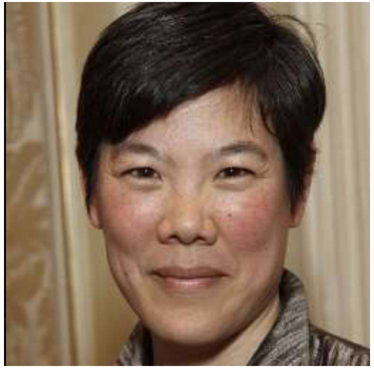
Spot the deepfake. There is exactly one deepfake.



A



B



C

© 2022 Carnegie Mellon University

[[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution

What is a deepfake?

“Believable media generated by a deep neural network”

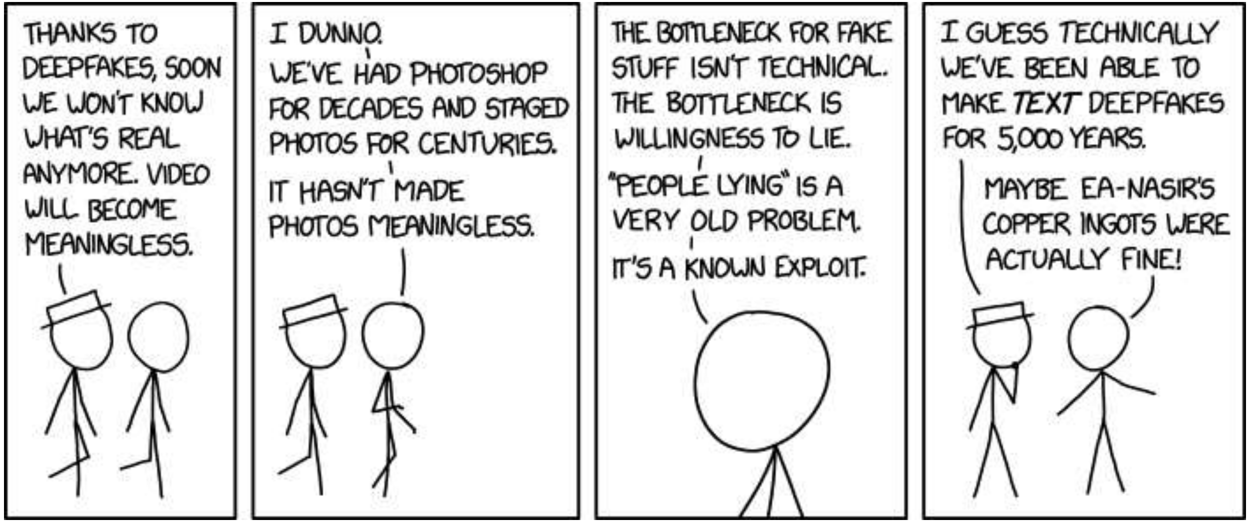
-- Mirsky and Lee (2020)

© 2022 Carnegie Mellon University

[[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution

Why should you care?

- Accessibility
- Automation
- Scalability

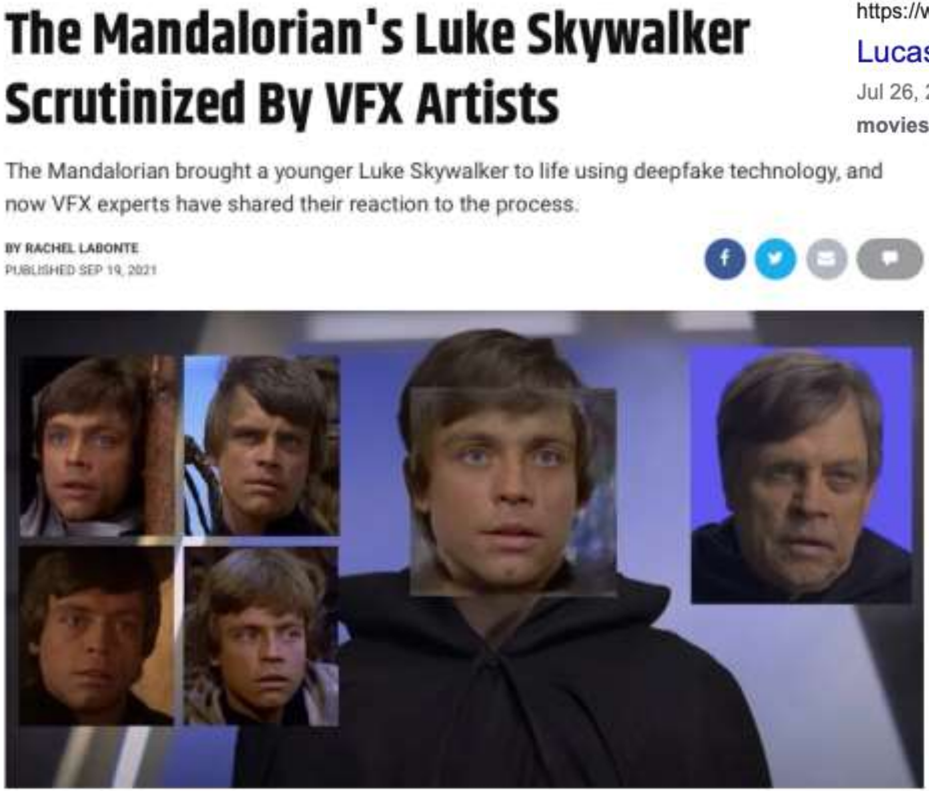

THANKS TO DEEPFAKES, SOON WE WON'T KNOW WHAT'S REAL ANYMORE. VIDEO WILL BECOME MEANINGLESS.
I DUNNO.
WE'VE HAD PHOTOSHOP FOR DECADES AND STAGED PHOTOS FOR CENTURIES.
IT HASN'T MADE PHOTOS MEANINGLESS.
THE BOTTLENECK FOR FAKE STUFF ISN'T TECHNICAL. THE BOTTLENECK IS WILLINGNESS TO LIE.
"PEOPLE LYING" IS A VERY OLD PROBLEM.
IT'S A KNOWN EXPLOIT.
I GUESS TECHNICALLY WE'VE BEEN ABLE TO MAKE TEXT DEEPFAKES FOR 5,000 YEARS.
MAYBE EA-NASIR'S COPPER INGOTS WERE ACTUALLY FINE!

From xkcd.com (#2650)

Deepfakes in Live Action

nthony Bourdain V

[illegible]



The Mandalorian's Luke Skywalker Scrutinized By VFX Artists

The Mandalorian brought a younger Luke Skywalker to life using deepfake technology, and now VFX experts have shared their reaction to the process.

BY RACHEL LABONTE
PUBLISHED SEP 19, 2021

https://www.indiewire.com › 2021/07 › lucasfilm-hires-...

Lucasfilm Hires Deepfake YouTuber Who Fixed Mandalorian ...

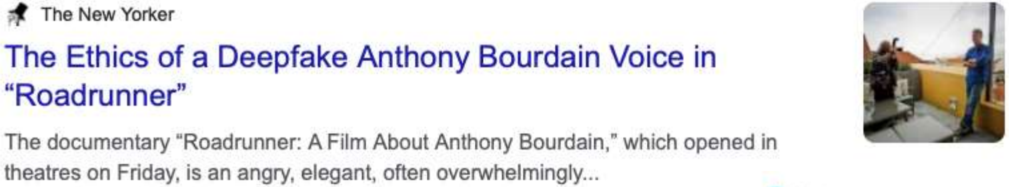
Jul 26, 2021 — Lucasfilm Hired the YouTuber **Who Used Deepfakes** to Tweak Luke ... effects in **movies**, but rarely do these videos lead to actual studio jobs.

https://gizmodo.com › deepfake-lips-are-coming-to-du...

Deepfake Lips Are Coming to Dubbed Films

May 6, 2021 — For those **who** find reading a foreign **film's** subtitles too distracting, **movies are** often dubbed into different languages.

Synthetic Voice Deepfakes

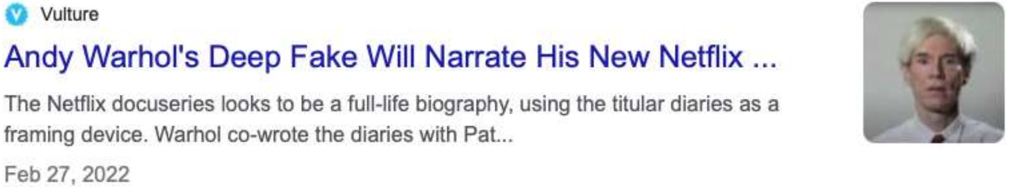


The New Yorker

The Ethics of a Deepfake Anthony Bourdain Voice in "Roadrunner"

The documentary "Roadrunner: A Film About Anthony Bourdain," which opened in theatres on Friday, is an angry, elegant, often overwhelmingly...

Jul 17, 2021

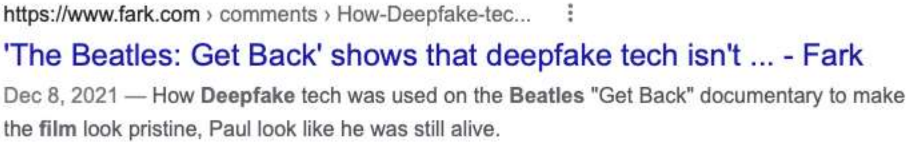


Vulture

Andy Warhol's Deep Fake Will Narrate His New Netflix ...

The Netflix docuseries looks to be a full-life biography, using the titular diaries as a framing device. Warhol co-wrote the diaries with Pat...

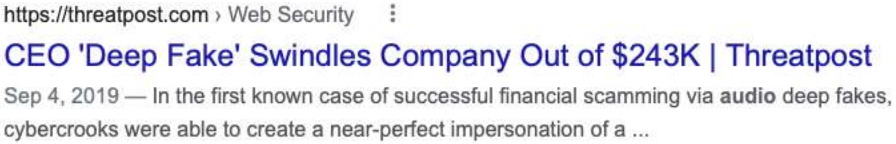
Feb 27, 2022



https://www.fark.com › comments › How-Deepfake-tec...

'The Beatles: Get Back' shows that deepfake tech isn't ... - Fark

Dec 8, 2021 — How **Deepfake** tech was used on the **Beatles** "Get Back" documentary to make the **film** look pristine, Paul look like he was still alive.



https://threatpost.com › Web Security

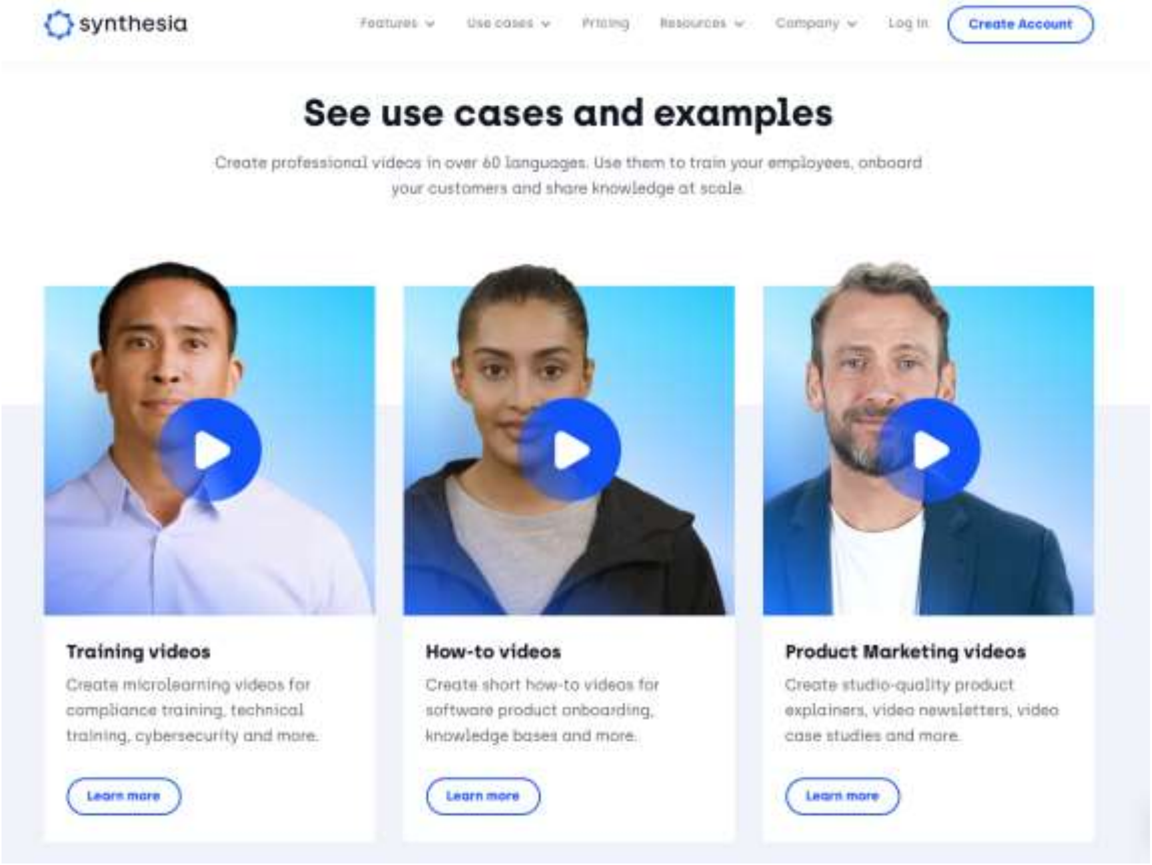
CEO 'Deep Fake' Swindles Company Out of $243K | Threatpost

Sep 4, 2019 — In the first known case of successful financial scamming via **audio** deep fakes, cybercrooks were able to create a near-perfect impersonation of a ...

Synthetic Avatars

nthony Bourdain V

[illegible]



See use cases and examples

Create professional videos in over 60 languages. Use them to train your employees, onboard your customers and share knowledge at scale.

Training videos

Create microlearning videos for compliance training, technical training, cybersecurity and more.

How-to videos

Create short how-to videos for software product onboarding, knowledge bases and more.

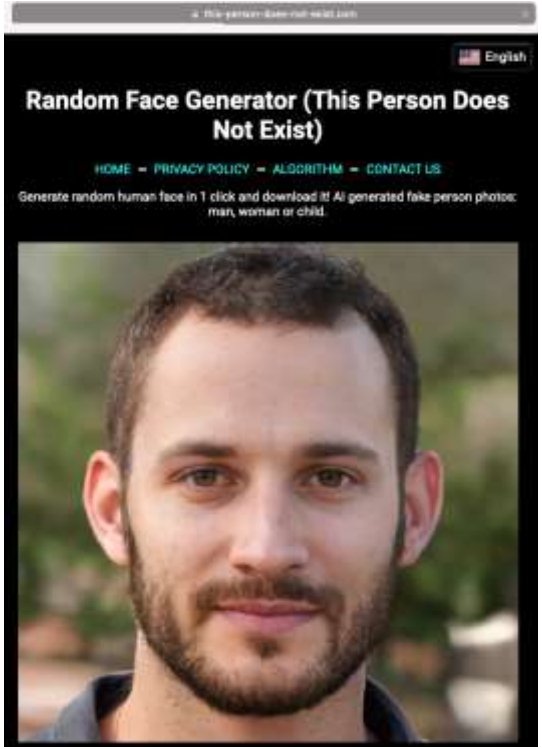
Product Marketing videos

Create studio-quality product explainers, video newsletters, video case studies and more.

nthony Bourdain V

[illegible]

This Person Does Not Exist



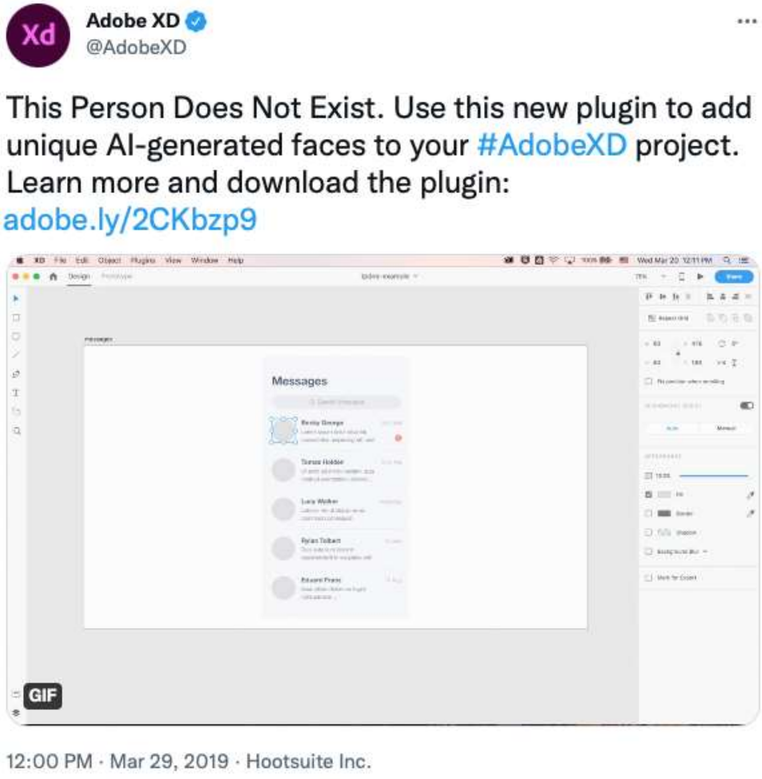
this-person-does-not-exist.com

English

Random Face Generator (This Person Does Not Exist)

HOME – PRIVACY POLICY – ALGORITHM – CONTACT US

Generate random human face in 1 click and download it! AI generated fake person photos: man, woman or child.



Adobe XD

@AdobeXD

This Person Does Not Exist. Use this new plugin to add unique AI-generated faces to your #AdobeXD project. Learn more and download the plugin: adobe.ly/2CKbzp9

12:00 PM · Mar 29, 2019 · Hootsuite Inc.

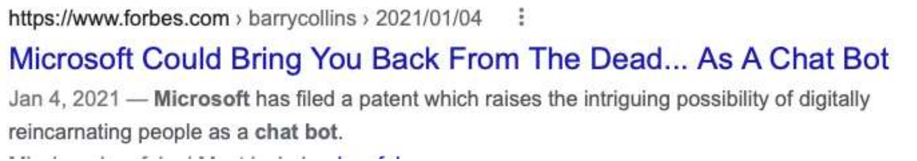
© 2022 Carnegie Mellon University

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution

Deepfakes in Depth

nthony Bourdain V

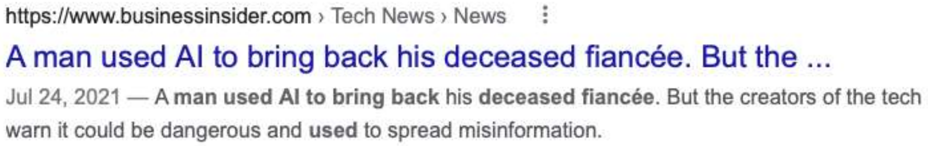
[illegible]



https://www.forbes.com › barrycollins › 2021/01/04

Microsoft Could Bring You Back From The Dead... As A Chat Bot

Jan 4, 2021 — **Microsoft** has filed a patent which raises the intriguing possibility of digitally reincarnating people as a **chat bot**.



https://www.businessinsider.com › Tech News › News

A man used AI to bring back his deceased fiancée. But the ...

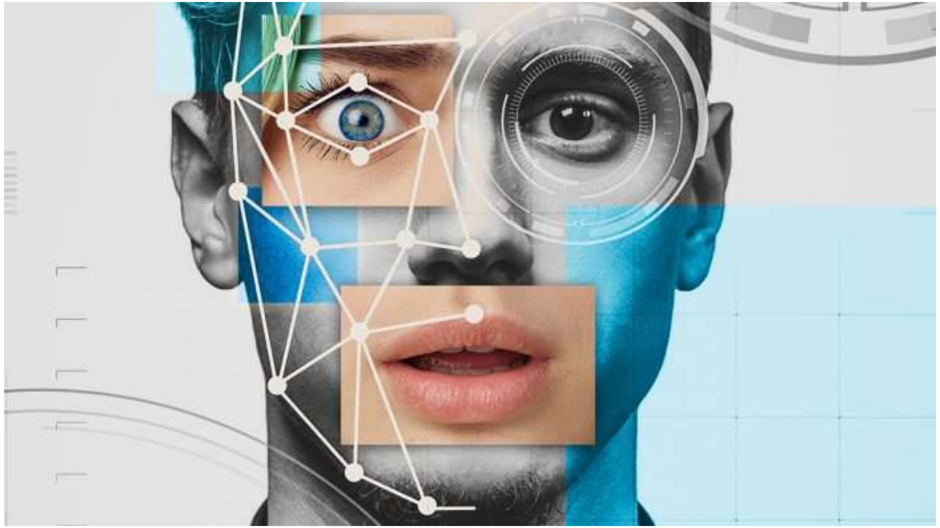
Jul 24, 2021 — A **man used AI to bring back** his **deceased fiancée**. But the creators of the tech warn it could be dangerous and **used** to spread misinformation.

How do we detect deepfakes?

Digital fingerprints

- Basic idea is humans leave evidence of their work (even in AI!)

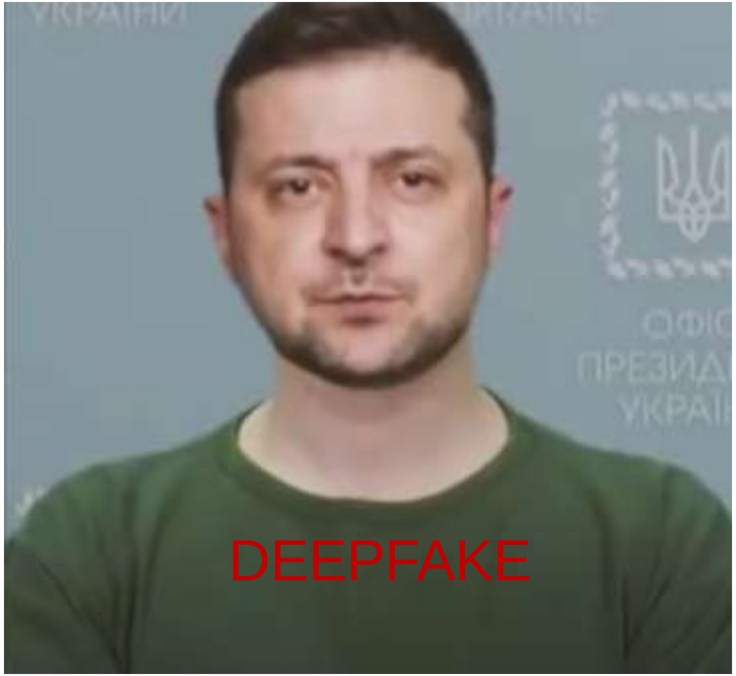
- Human extracted vs. computer extracted fingerprints



Deepfakes and how can we detect them?

Deepfake The Art vs the Algorithm

Post-processing – Why does this look bad?



The deepfake appeared on the hacked website of Ukrainian TV network Ukrayina 24

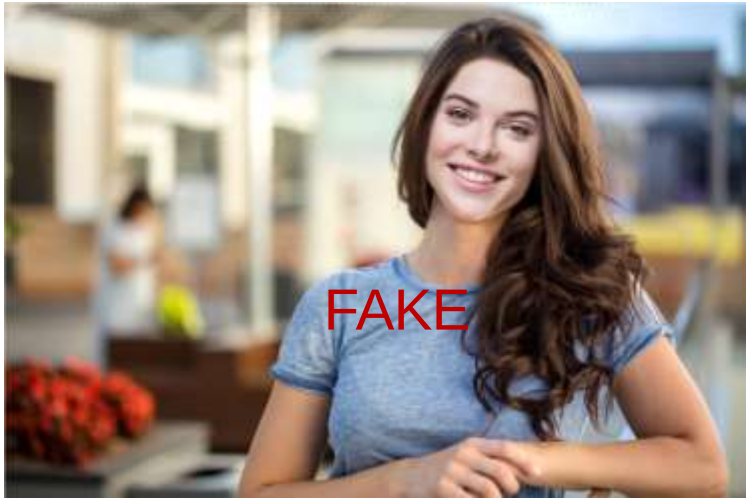
Post-processing – Not a Deepfake

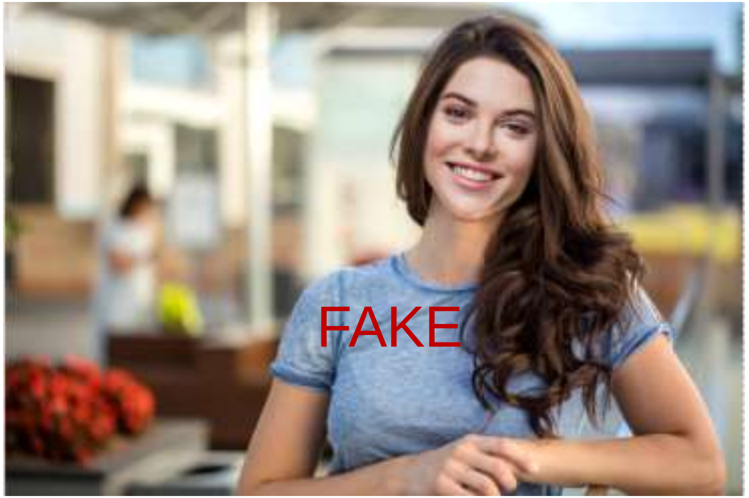
The examples shown are not deepfakes, and are not part of active SEI research. Images shown demonstrate post-processing and not meant to be a representation of deepfake generation.

© 2022 Carnegie Mellon University

[[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution

Post-processing – Typical Result



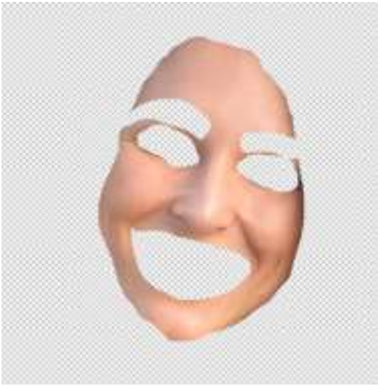


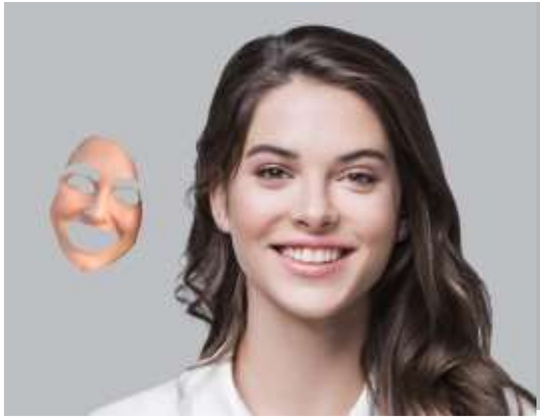
Example of Post Processing Techniques not made with a Neural Network.

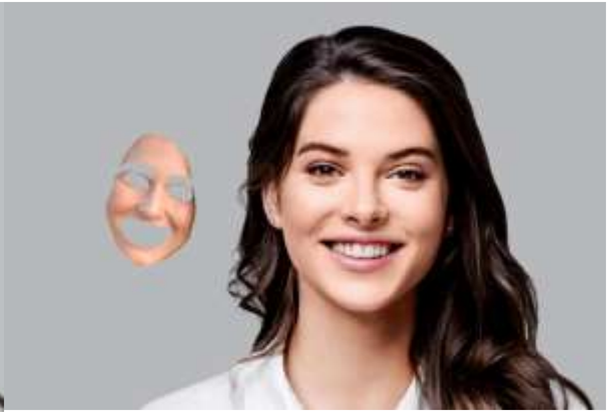
Post-processing – Isolation

Example of Post Processing Techniques not made with a Neural Network.

Post-processing – Color Correction



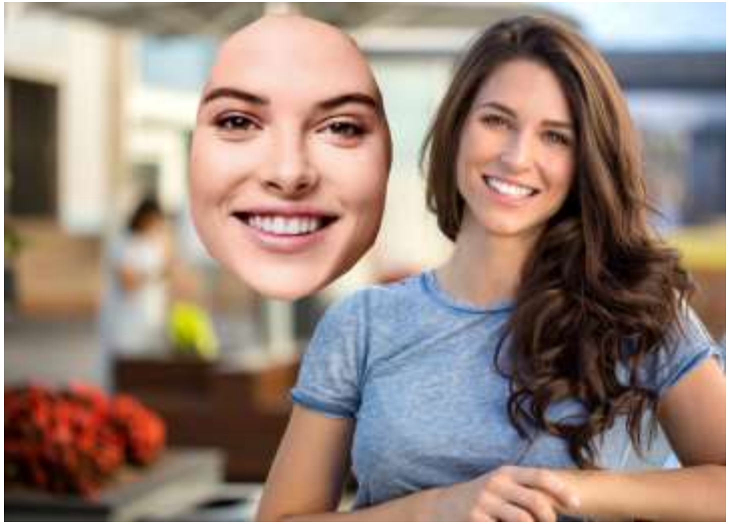


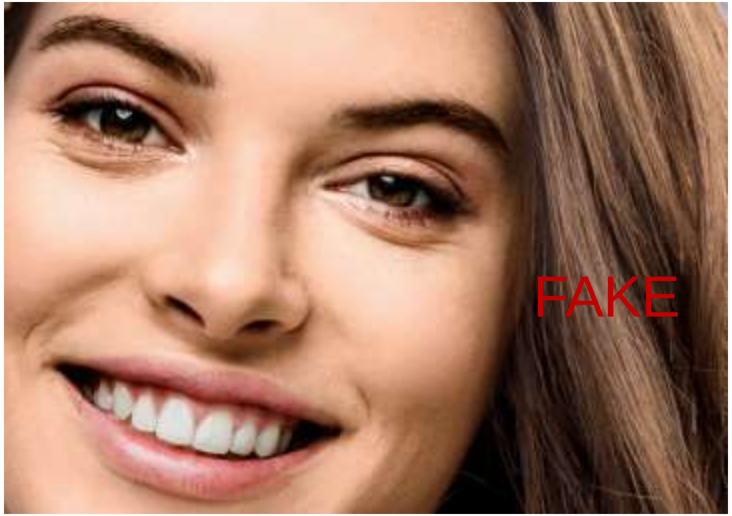


© 2022 Carnegie Mellon University

[[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution

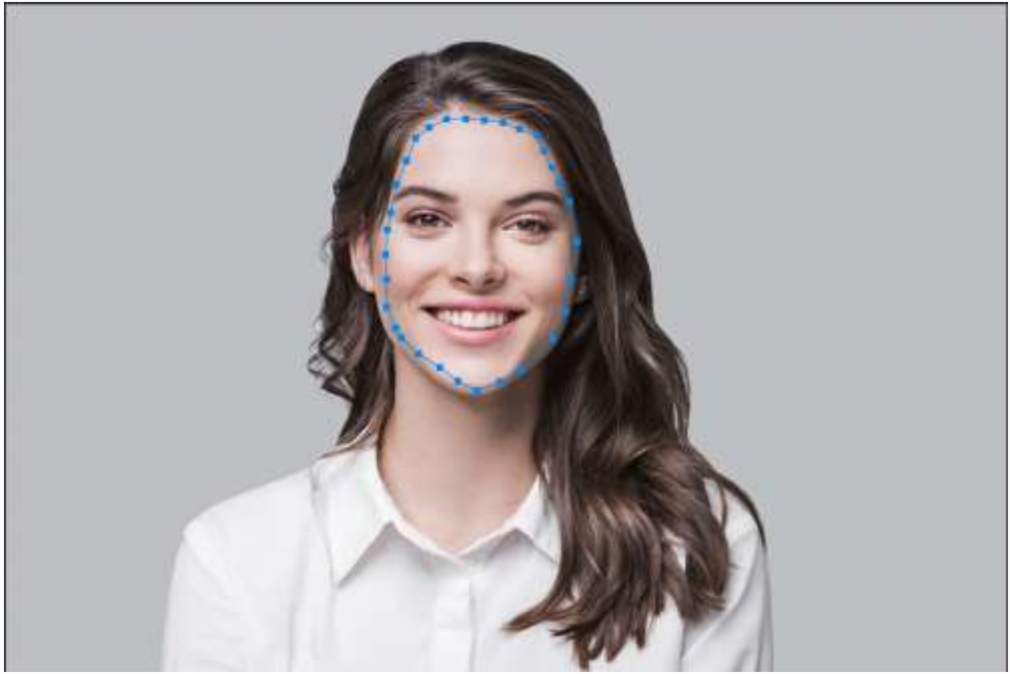
Post-processing – Color Correction





Example of Post Processing Techniques not made with a Neural Network.

Post-processing – Motion Tracking

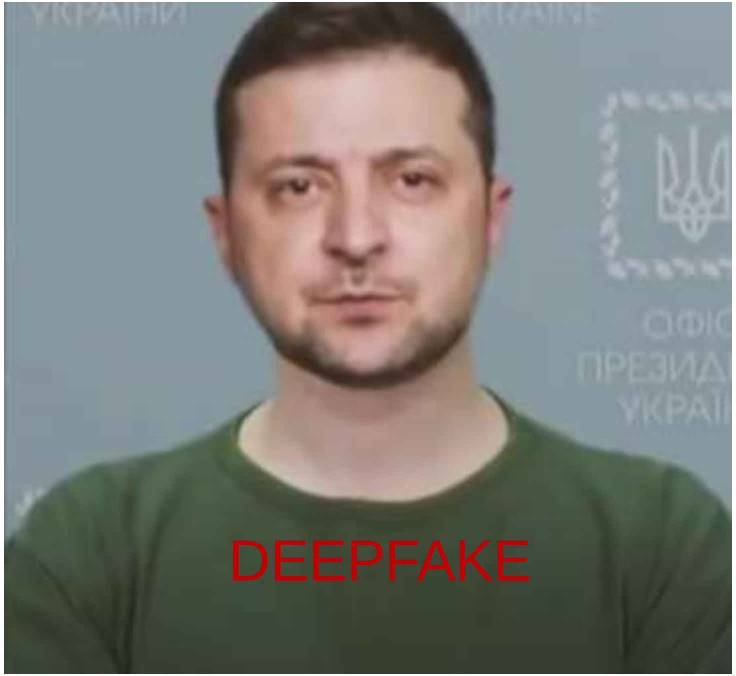


Example of Post Processing Techniques not made with a Neural Network.

Post-processing – Compositing

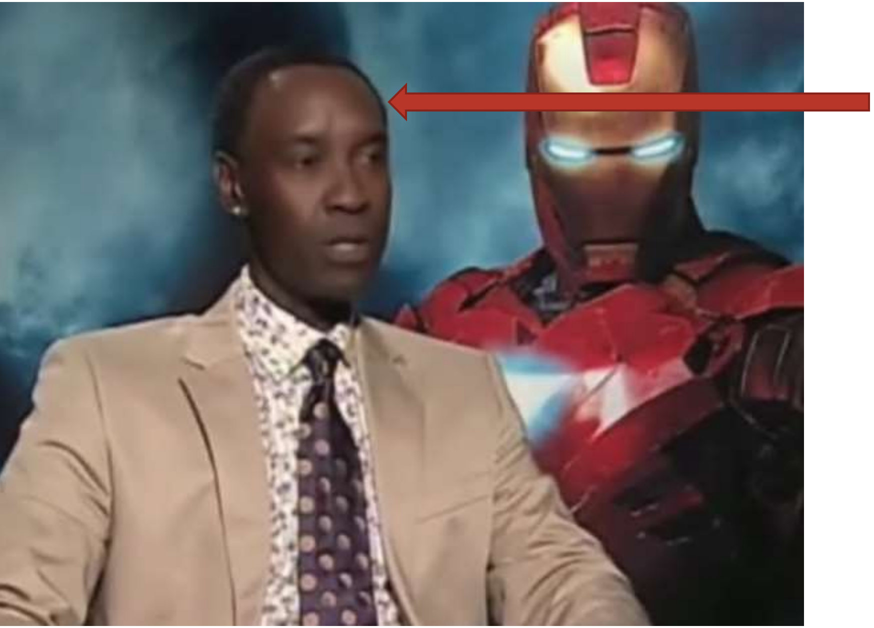
Example of Post Processing Techniques not made with a Neural Network.

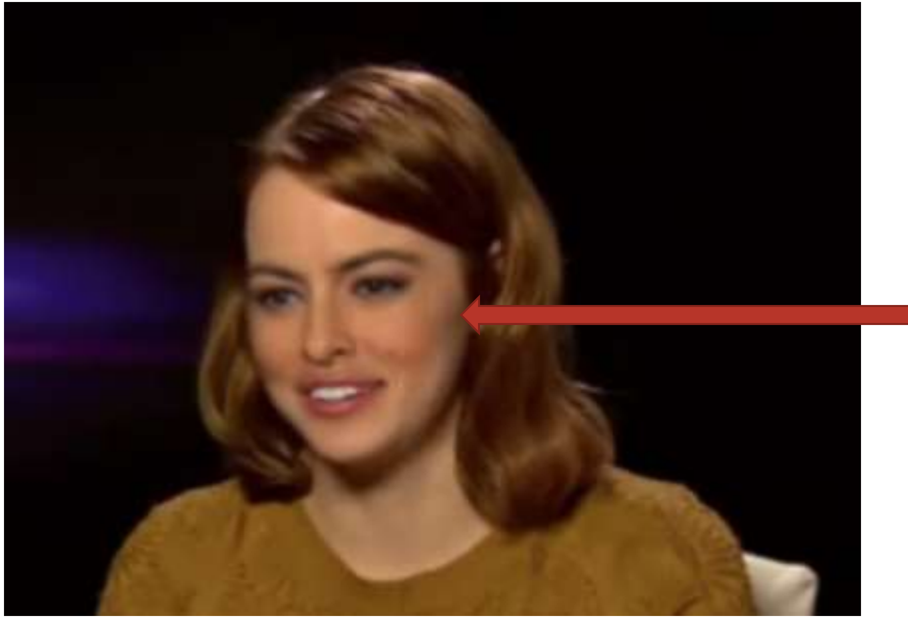
Not photorealistic for a reason.



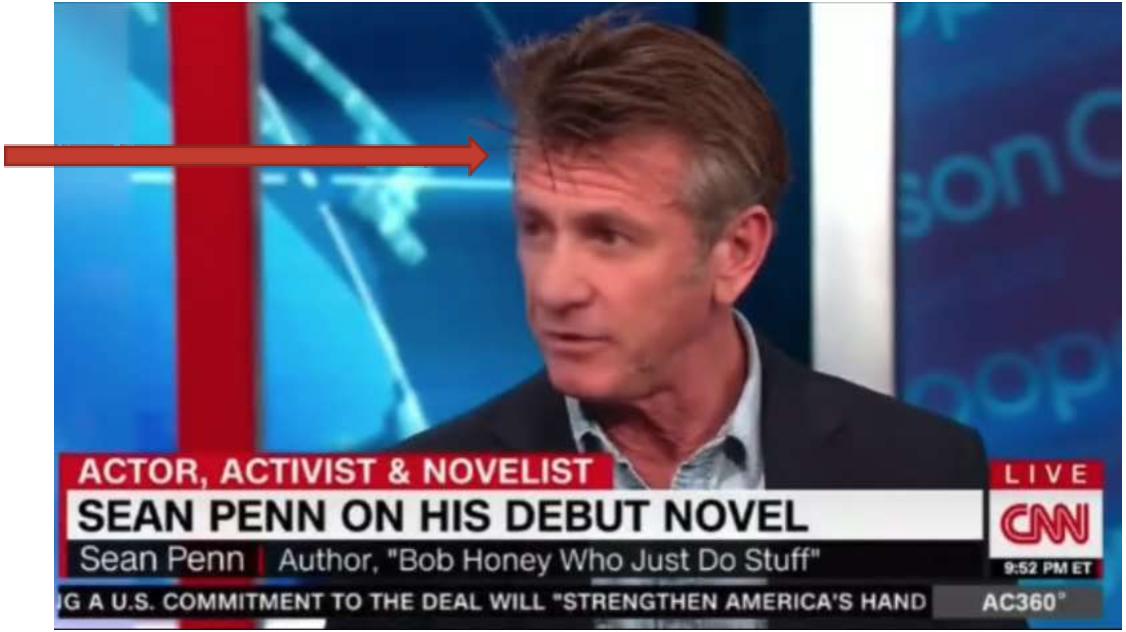
The deepfake appeared on the hacked website of Ukrainian TV network Ukrayina 24

Spot the Deepfake – Celeb DF-2 Dataset





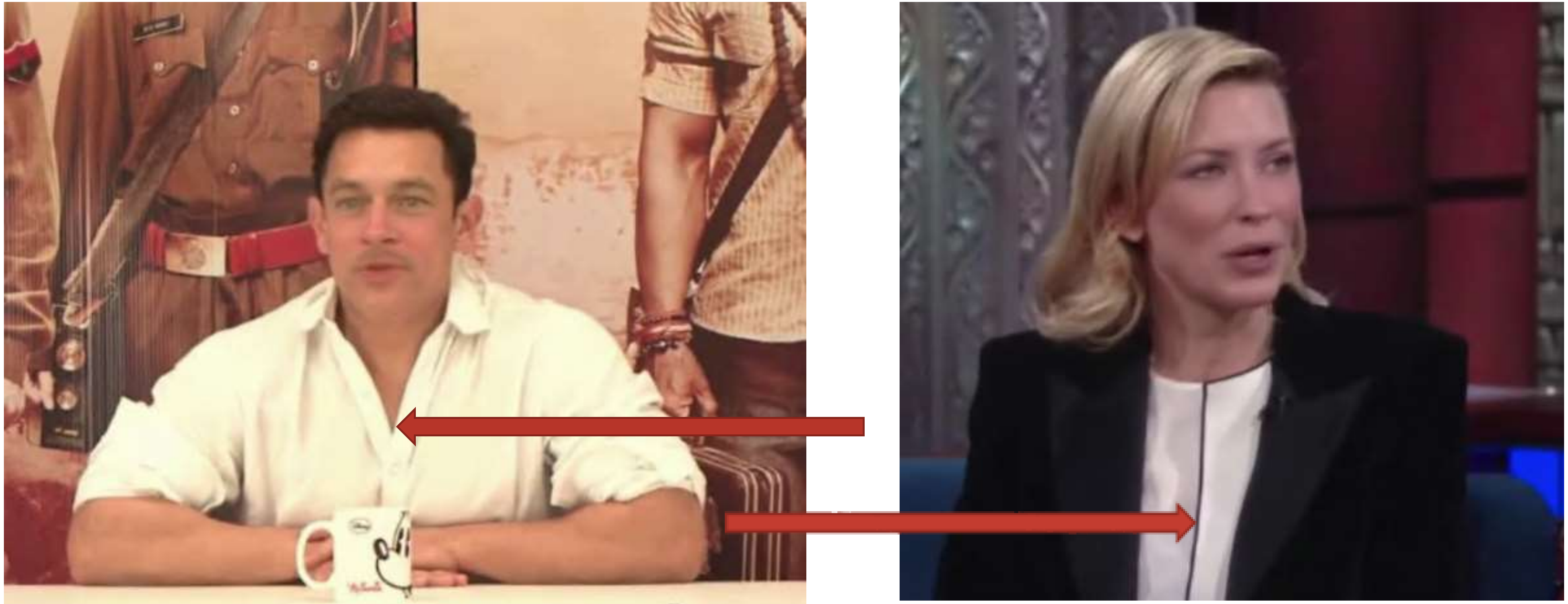
Spot the Deepfake – Celeb DF-2 Dataset


ACTOR, ACTIVIST & NOVELIST
SEAN PENN ON HIS DEBUT NOVEL
Sean Penn | Author, "Bob Honey Who Just Do Stuff"
LIVE
CNN
9:52 PM ET
G A U.S. COMMITMENT TO THE DEAL WILL "STRENGTHEN AMERICA'S HAND
AC360°

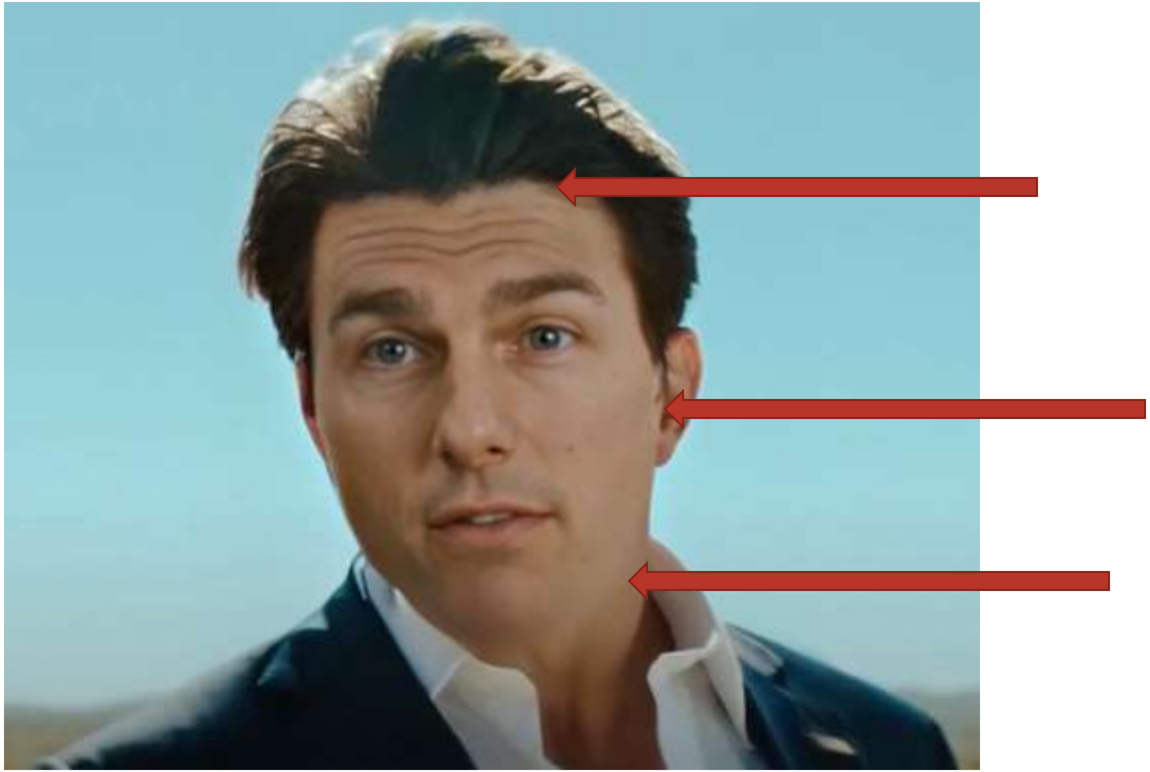
© 2022 Carnegie Mellon University

[[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution

Spot the Deepfake – Celeb DF-2 Dataset



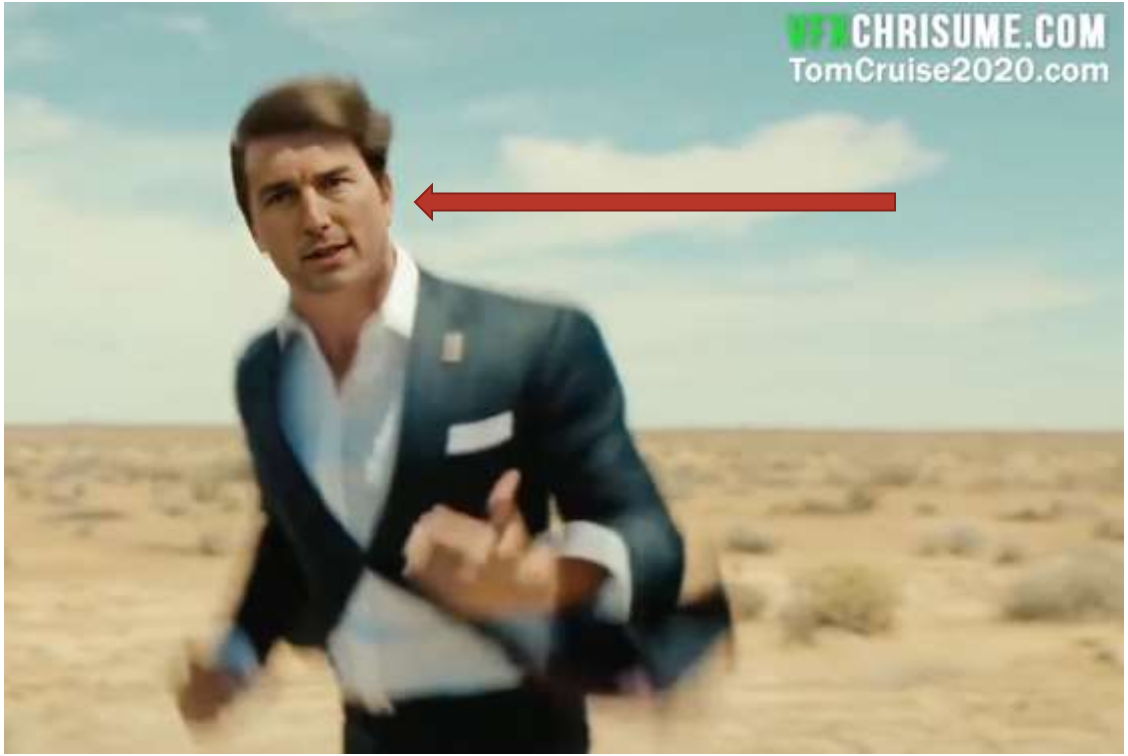
Spot the Deepfake – Chris Ume - Metaphysics AI



© 2022 Carnegie Mellon University

[[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution

Spot the Deepfake – Chris Ume - Metaphysics AI



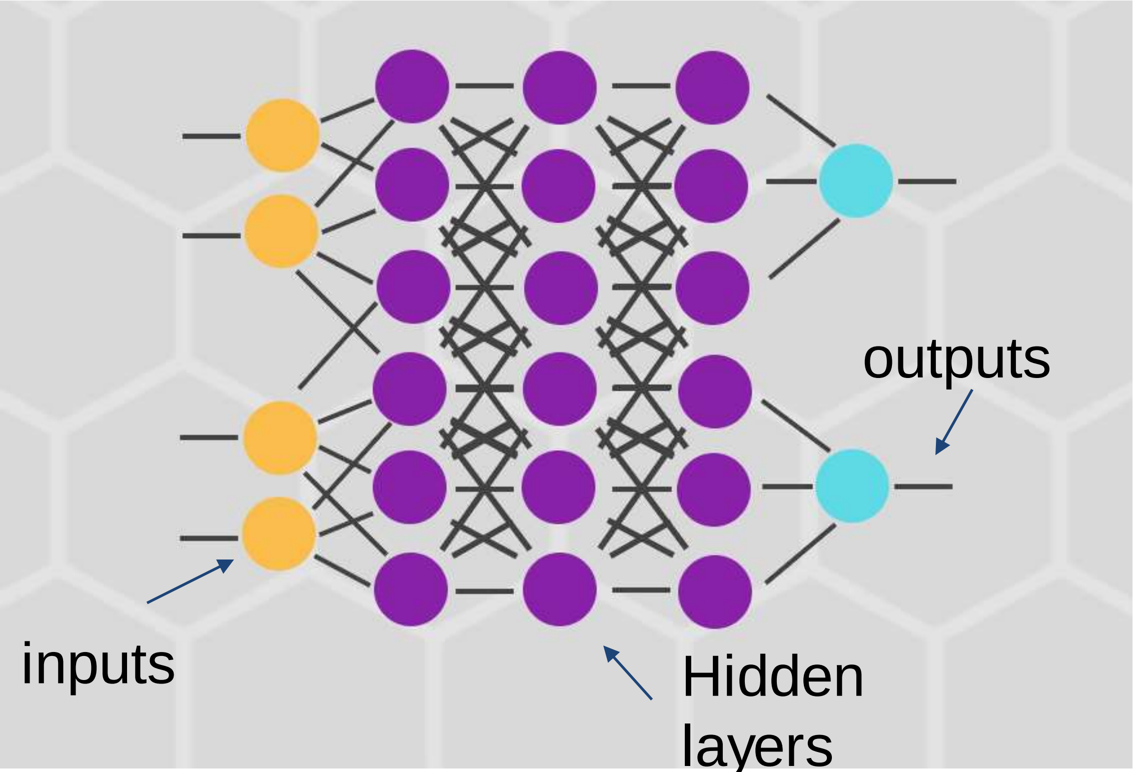
So we know what features we need to look for!

- **Blending**
- **Color**
- **Focus**
- **Discontinuity**
- **Blinking**

With >500k hours of video uploaded everyday, cannot rely on expert human detection

Deep neural network detectors

- Inputs of dimension (W, H, C, F)
  - W = Pixel width
  - H = Pixel height
  - C = Channel (RGB)
  - F = Frame #
- Hidden layers are where machine extracts digital fingerprints
- Outputs: real vs. fake



This Photo by Unknown Author is licensed under CC BY

Neural nets can do great when testing on in-distribution data, but suffer greatly when data distribution ‘drifts’

Accuracy (%) of fine-tuned ResNet

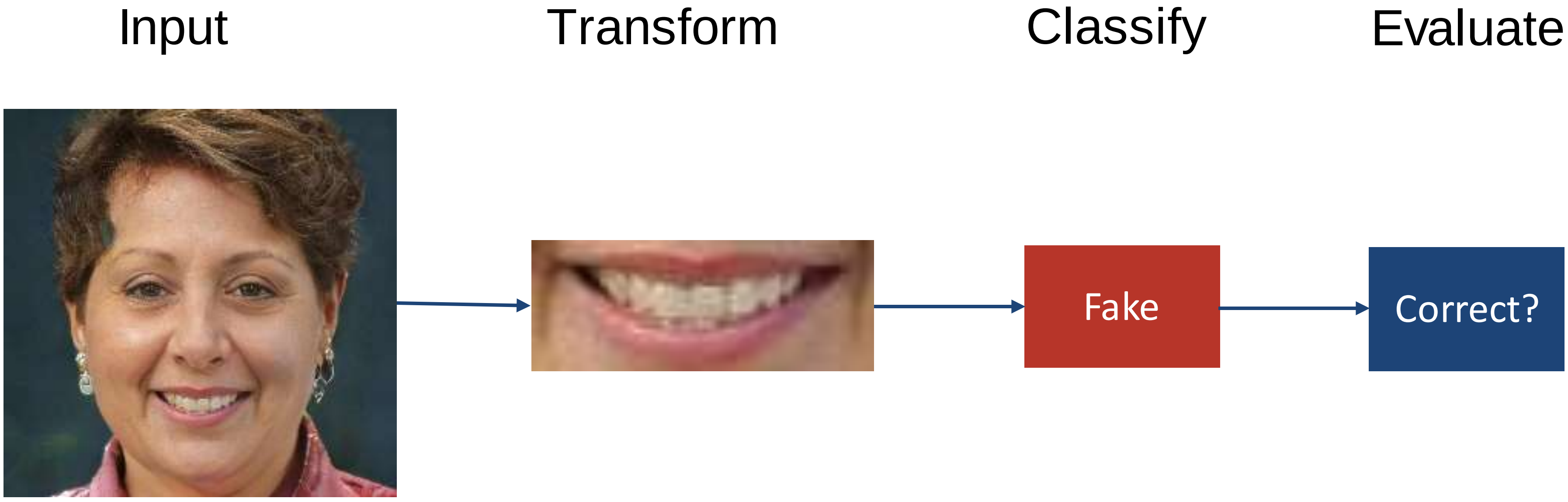
Tested on

	Data Set	Celeb DF v1	Stylegan2	Stylegan3-t	Stylegan3-r	DFDC Pt. 0
Trained on	**Celeb DF v1**	99.1	44.2	44.2	44.0	51.2
	Stylegan2	24.1	98.7	52.9	48.4	57.4
	Stylegan3-t	16.7	69.7	96.7	84.0	7.0
	Stylegan3-r	16.9	68.0	89.0	97.2	7.0
	DFDC Pt. 0	68.1	57.4	57.5	57.5	88.7

© 2022 Carnegie Mellon University

[[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution

We need a pipeline! E.g.



Input

Transform

Classify

Evaluate

Fake

Correct?

Image from Stylegan2

False Positive vs. False Negative

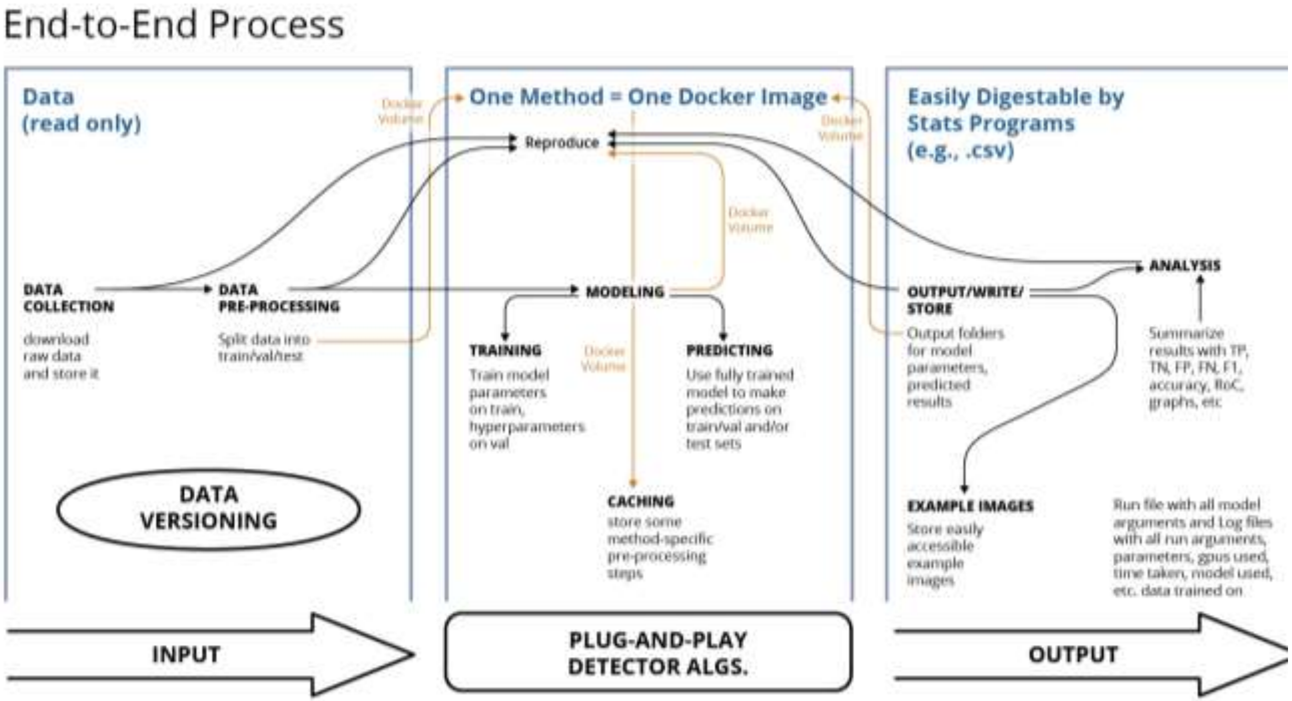
		Classified As	
		FAKE	**REAL**
Actual Label	**FAKE**	**True Positive**	**False Negative**
	REAL	**False Positive**	**True Negative**

- False Pos. = real image classified as fake
- False Neg. = fake image classified as real

© 2022 Carnegie Mellon University

[[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution

Our approach: Deepfake Detection Pipeline (DDP)



End-to-End Process

Data (read only)

DATA COLLECTION

download raw data and store it

DATA PRE-PROCESSING

Split data into train/val/test

DATA VERSIONING

INPUT

One Method = One Docker Image

Reproduce

MODELING

TRAINING

Train model parameters on train, hyperparameters on val

PREDICTING

Use fully trained model to make predictions on train/val and/or test sets

CACHING

store some method-specific pre-processing steps

PLUG-AND-PLAY DETECTOR ALGS.

Easily Digestable by Stats Programs (e.g., .csv)

OUTPUT/WRITE/STORE

Output folders for model parameters, predicted results

ANALYSIS

Summarize results with TP, TN, FP, FN, F1, accuracy, RoC, graphs, etc

EXAMPLE IMAGES

Store easily accessible example images

Run file with all model arguments and Log files with all run arguments, parameters, gpus used, time taken, model used, etc. data trained on

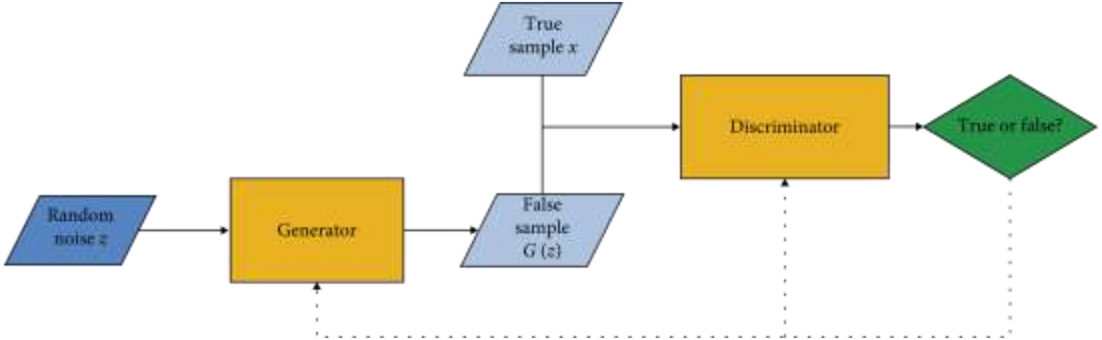
OUTPUT

DDP is reproducible, portable, and modular

DDP’s backend is SEI’s Juneberry – a publicly accessible tool!

The ever-present threat: generative adversarial models

- **Adversaries can use our detector results** to make improved images that can better fool the detector



This Photo by Unknown Author is licensed under CC BY

© 2022 Carnegie Mellon University
[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution

In closing

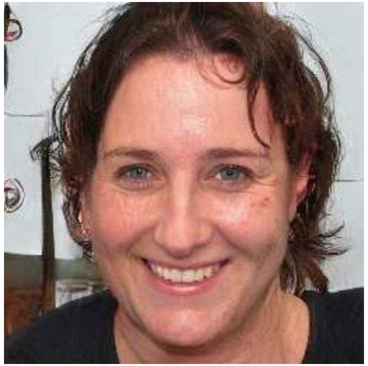
- **Deepfakes are modern lies**
    - Scalable and automated

- **They are an art**

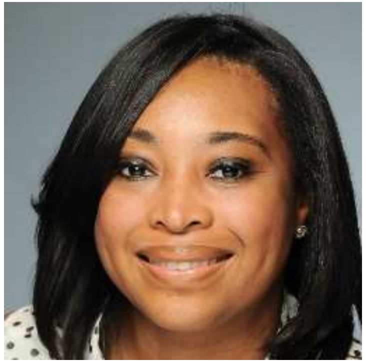
- **But can benefit greatly from computer assistance**

- **Better detectors can directly lead to better generators, leading to a multi-round game**

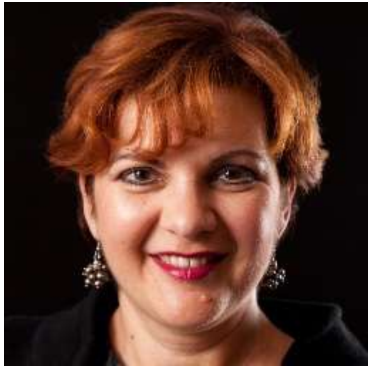
Spot the deepfake. There is exactly one deepfake.



A



B



C

© 2022 Carnegie Mellon University

[[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution