TRUSTED VISION AI

PARAVISION

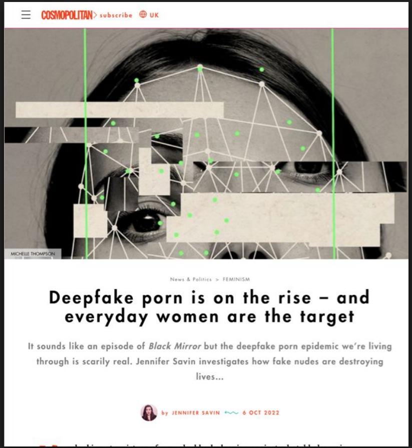
Synthetics and Deepfakes

Opportunities, Threats, and Protection

Neda Eskandari, Machine Learning Lead at Paravision

This presentation material contains information which is proprietary to Paravision. Information contained herein shall not be disclosed or shared with others, copied, reproduced or otherwise used in whole or in part without written consent.

NIST IFPC 2022

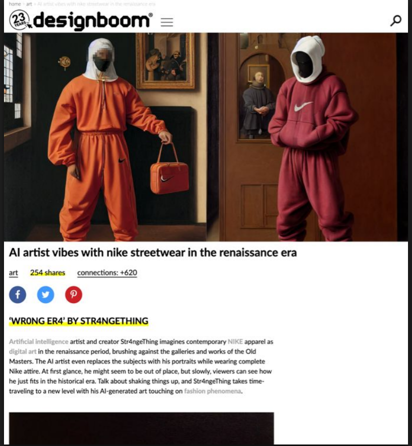


News & Politics > FEMINISM

Deepfake porn is on the rise – and everyday women are the target

It sounds like an episode of *Black Mirror* but the deepfake porn epidemic we're living through is scarily real. Jennifer Savin investigates how fake nudes are destroying lives...

by JENNIFER SAVIN 6 OCT 2022

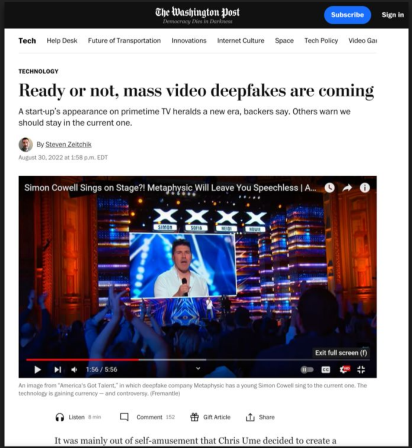


AI artist vibes with nike streetwear in the renaissance era

art 254 shares connections: +620

'WR0NG ER4' BY STR4NGETHING

Artificial intelligence artist and creator Str4ngeThing imagines contemporary NIKE apparel as digital art in the renaissance period, brushing against the galleries and works of the Old Masters. The AI artist even replaces the subjects with his portraits while wearing complete Nike attire. At first glance, he might seem to be out of place, but slowly, viewers can see how he just fits in the historical era. Talk about shaking things up, and Str4ngeThing takes time-traveling to a new level with his AI-generated art touching on fashion phenomena.



TECHNOLOGY

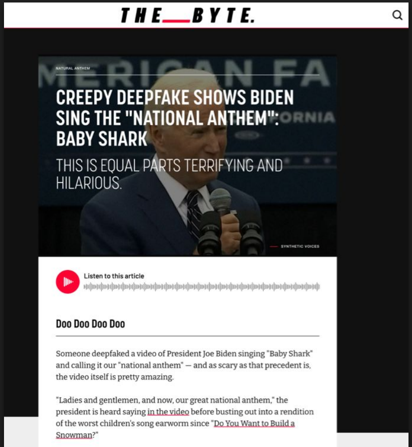
Ready or not, mass video deepfakes are coming

A start-up's appearance on primetime TV heralds a new era, backers say. Others warn we should stay in the current one.

By Steven Zeitchik
August 30, 2022 at 1:58 p.m. EDT

An image from "America's Got Talent," in which deepfake company Metaphysic has a young Simon Cowell sing to the current one. The technology is gaining currency — and controversy. (Fremantle)

It was mainly out of self-amusement that Chris Ume decided to create a



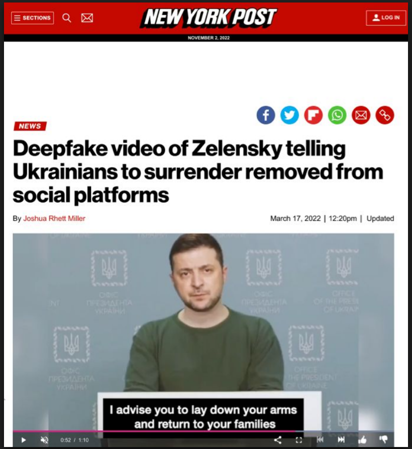
CREEPY DEEPFAKE SHOWS BIDEN SING THE "NATIONAL ANTHEM": BABY SHARK

THIS IS EQUAL PARTS TERRIFYING AND HILARIOUS.

Doo Doo Doo Doo

Someone deepfaked a video of President Joe Biden singing "Baby Shark" and calling it our "national anthem" — and as scary as that precedent is, the video itself is pretty amazing.

"Ladies and gentlemen, and now, our great national anthem," the president is heard saying in the video before busting out into a rendition of the worst children's song earworm since "Do You Want to Build a Snowman?"

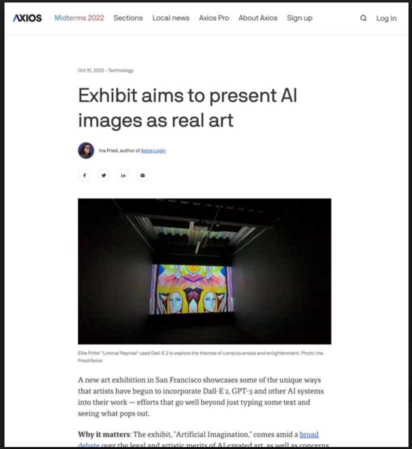


NEWS

Deepfake video of Zelensky telling Ukrainians to surrender removed from social platforms

By Joshua Rhett Miller March 17, 2022 | 12:20pm | Updated

I advise you to lay down your arms and return to your families



Oct 31, 2022 - Technology

Exhibit aims to present AI images as real art

Ina Fried, author of Axios Login

Ellie Pritts' "Liminal Reprise" used Dall-E 2 to explore the themes of consciousness and enlightenment. Photo: Ina Fried/Axios

A new art exhibition in San Francisco showcases some of the unique ways that artists have begun to incorporate Dall-E 2, GPT-3 and other AI systems into their work — efforts that go well beyond just typing some text and seeing what pops out.

Why it matters: The exhibit, "Artificial Imagination," comes amid a broad debate over the legal and artistic merits of AI-created art, as well as concerns

TRUSTED **VISION AI**

History of deepfakes and synthetics

Before 2014

- → VAEs were available as generative models mainly for image retrieval

2014

- → GAN was introduced

2017-2018

- → Term “Deepfake” was introduced
- → Results of generative models became visually convincing

2021

- → Diffusion models were introduced
- → Report showed that the amount of deepfakes doubles every 6 months.

Deepfakes/Synthetic media in computer vision aren’t just one thing

01 Type of **data**

→ Video
→ Image

02 Scope of **imaging**

→ Face
→ Body
→ Background
→ Objects

03 Common **digital manipulations**

→ Identity swap
→ Expression swap
→ Audio to video
→ Text to video
→ Entire face synthesis
→ Face morphing
→ Attribute morphing
→ Adversarial attacks

Agenda

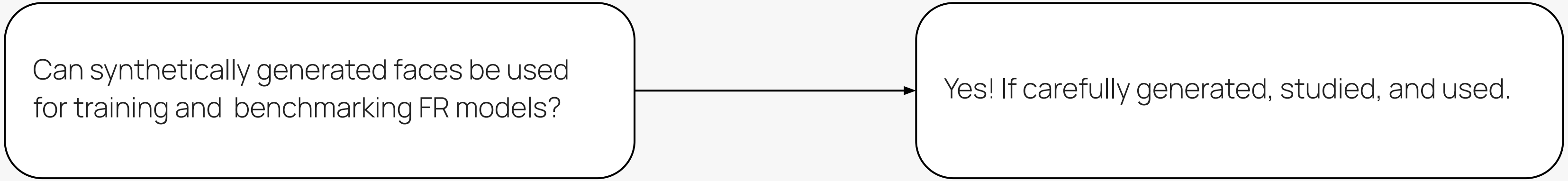
Opportunities

Threats

Protections

Opportunities

Opportunities: Entire face synthesis



Can synthetically generated faces be used for training and benchmarking FR models?

Yes! If carefully generated, studied, and used.

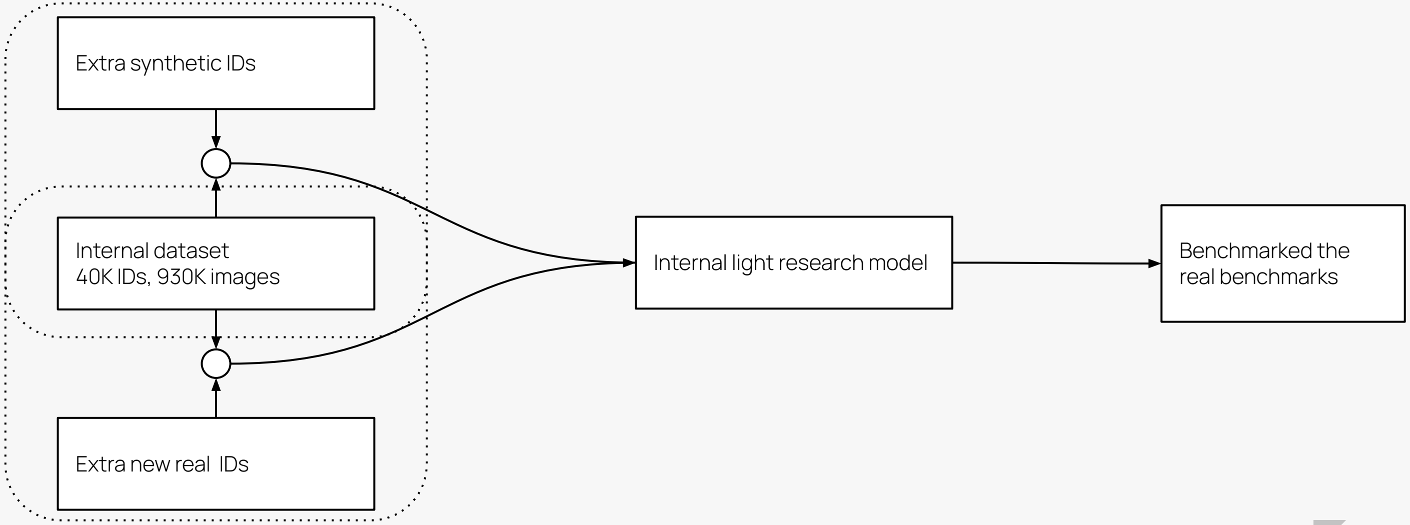
→ For expanding the training datasets: the synthetic characteristics of the data should not conceal the core characteristics of the real dataset.

→ For expanding the benchmarking datasets: synthetic embeddings have to be good representative of real embeddings.

TRUSTED VISION AI

Opportunities: Entire face synthesis

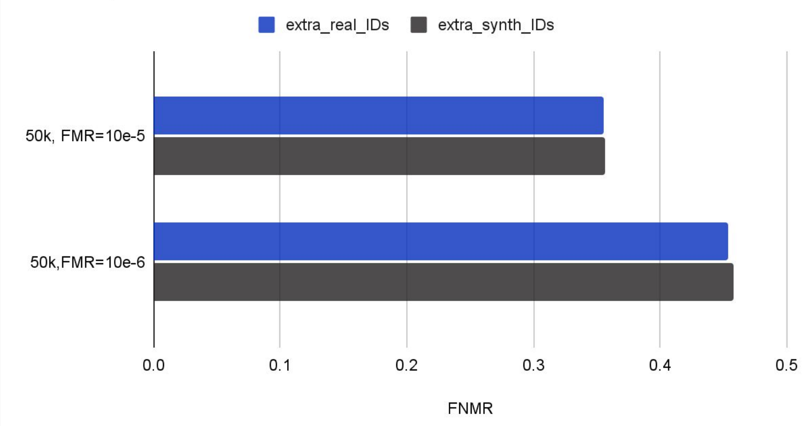
Training dataset expansion:



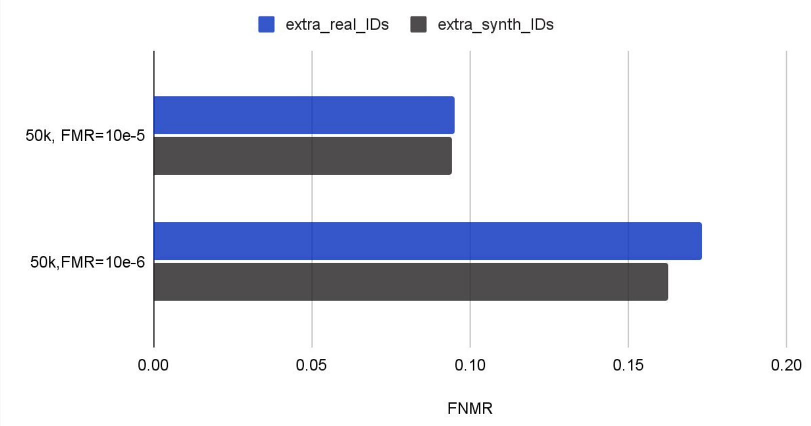
Opportunities: Entire face synthesis

Benchmarked real datasets:

Wild benchmark



Visa benchmark



Opportunities: Entire face synthesis

Extra real faces

→ Blurrier

→ More extreme attributes and poses

→ Various lightings

Extra synthetic faces

→ Higher quality

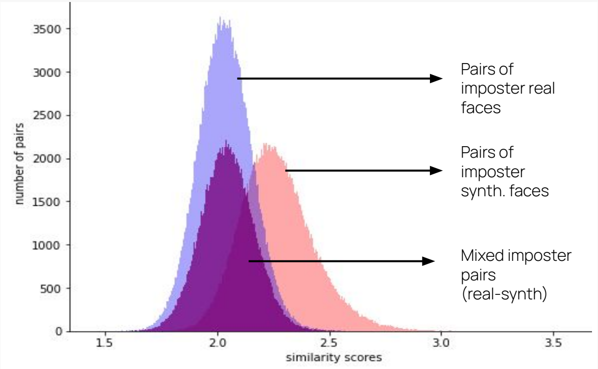
→ Less extreme poses

Opportunities: Entire face synthesis

Can synthetically generated faces be used for benchmarking FR models?

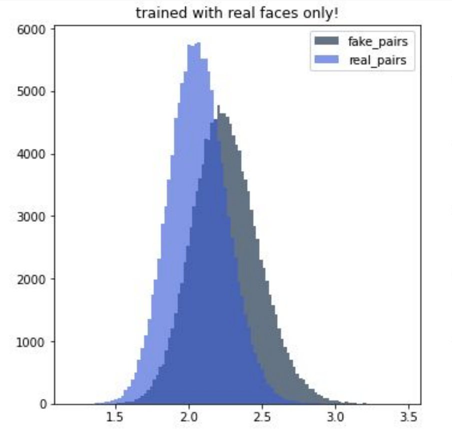
Not for the highly precise FR model that had never seen synthetically generated faces before.

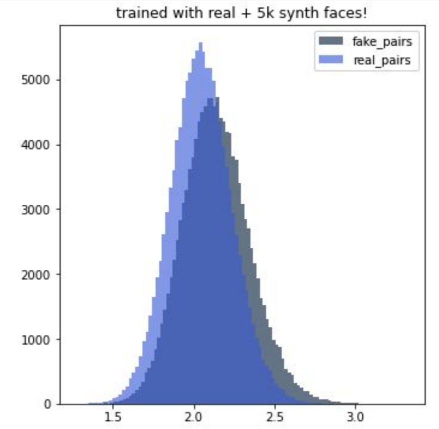
SOLUTION: Introducing synthetic faces to the model during training.

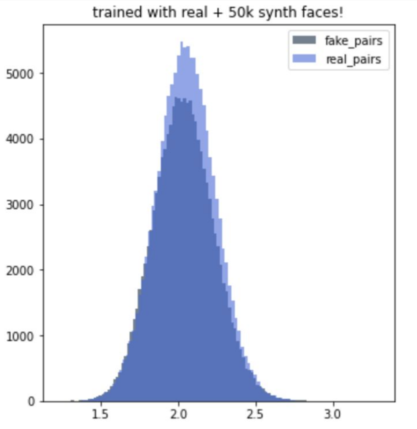


Opportunities: Entire face synthesis

Similarity distributions of real and synthetic face pairs:


trained with real faces only!


trained with real + 5k synth faces!


trained with real + 50k synth faces!

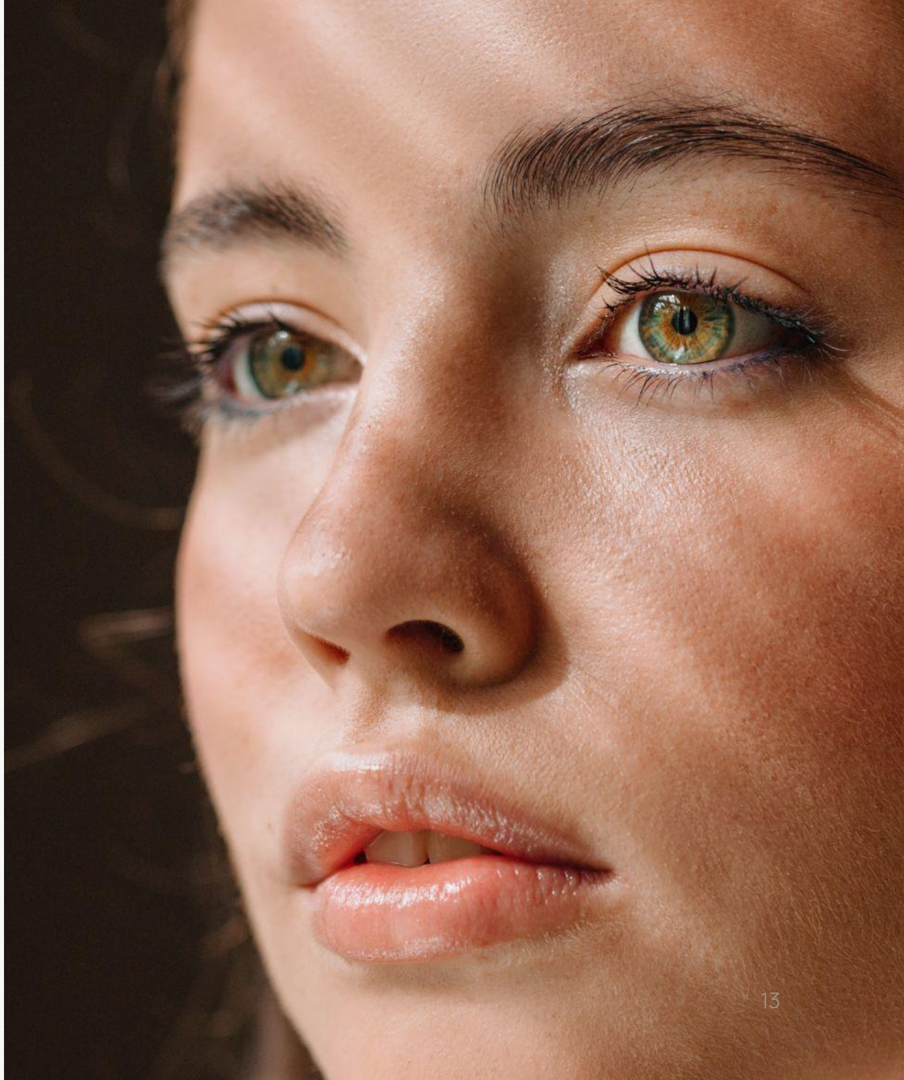
More synthetic faces seen by FR model during training

How to leverage entire face synthesis?

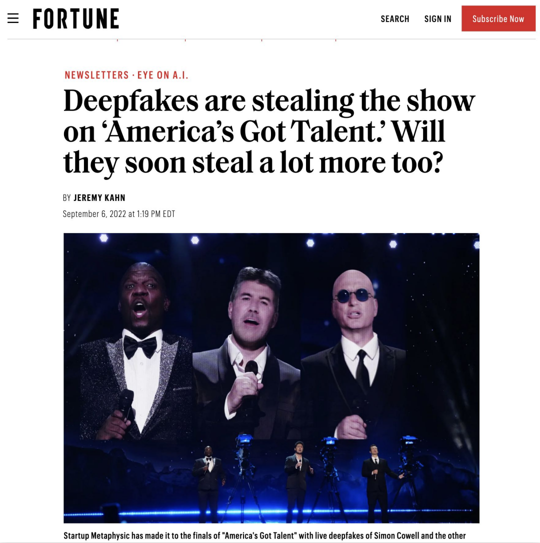
→ More careful design and selection of synth faces

→ Careful selection of synthetic/real ratio

→ Introducing multiple synthetic faces per ID



Opportunities: Deepfake videos



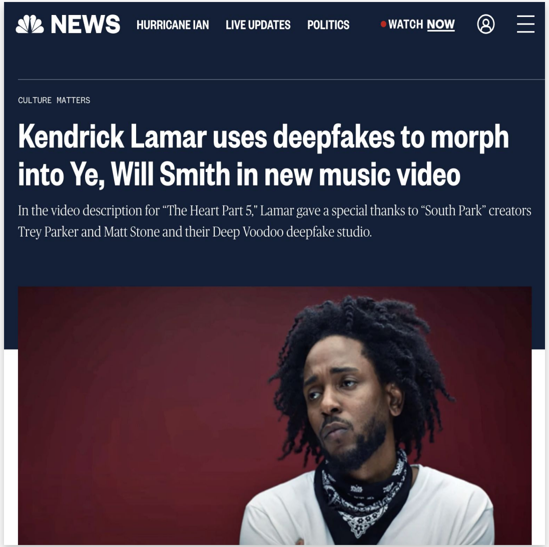
NEWSLETTERS · EYE ON A.I.

Deepfakes are stealing the show on ‘America’s Got Talent.’ Will they soon steal a lot more too?

BY JEREMY KAHN

September 6, 2022 at 1:19 PM EDT

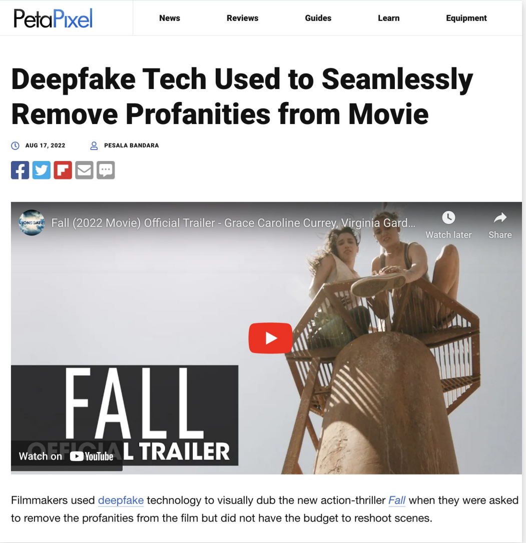
Startup Metaphysic has made it to the finals of "America's Got Talent" with live deepfakes of Simon Cowell and the other



CULTURE MATTERS

Kendrick Lamar uses deepfakes to morph into Ye, Will Smith in new music video

In the video description for “The Heart Part 5,” Lamar gave a special thanks to “South Park” creators Trey Parker and Matt Stone and their Deep Voodoo deepfake studio.



Deepfake Tech Used to Seamlessly Remove Profanities from Movie

AUG 17, 2022 PESALA BANDARA

Filmmakers used deepfake technology to visually dub the new action-thriller *Fall* when they were asked to remove the profanities from the film but did not have the budget to reshoot scenes.

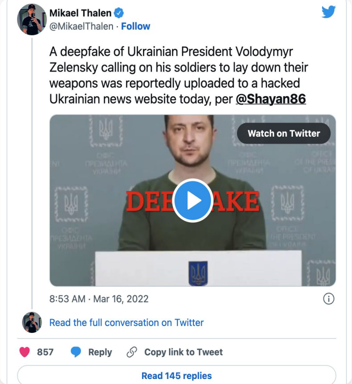
Threats

Deepfakes are an emerging threat to security

The technology to create hyper-realistic Deepfake imagery is now widely available.

Deepfakes have a potential to create significant risks for:

→ democracy
→ national security
→ business
→ human rights
→ personal privacy

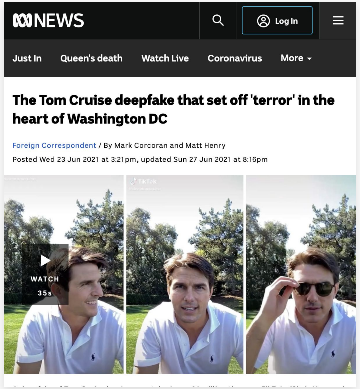


Mikael Thalen @MikaelThalen · Follow

A deepfake of Ukrainian President Volodymyr Zelensky calling on his soldiers to lay down their weapons was reportedly uploaded to a hacked Ukrainian news website today, per @Shayan86

8:53 AM · Mar 16, 2022

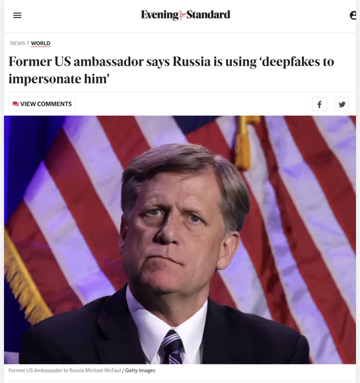
Read the full conversation on Twitter



The Tom Cruise deepfake that set off 'terror' in the heart of Washington DC

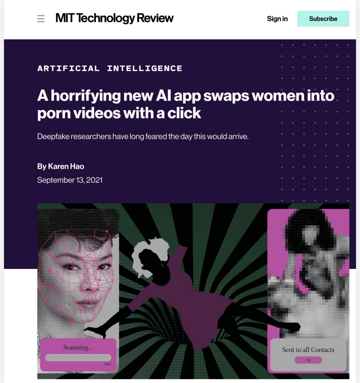
Foreign Correspondent / By Mark Corcoran and Matt Henry

Posted Wed 23 Jun 2021 at 3:21pm, updated Sun 27 Jun 2021 at 8:16pm



Former US ambassador says Russia is using 'deepfakes to impersonate him'

Former US Ambassador to Russia Michael McFaul / Getty Images



ARTIFICIAL INTELLIGENCE

A horrifying new AI app swaps women into porn videos with a click

Deepfake researchers have long feared the day this would arrive.

By Karen Hao
September 13, 2021

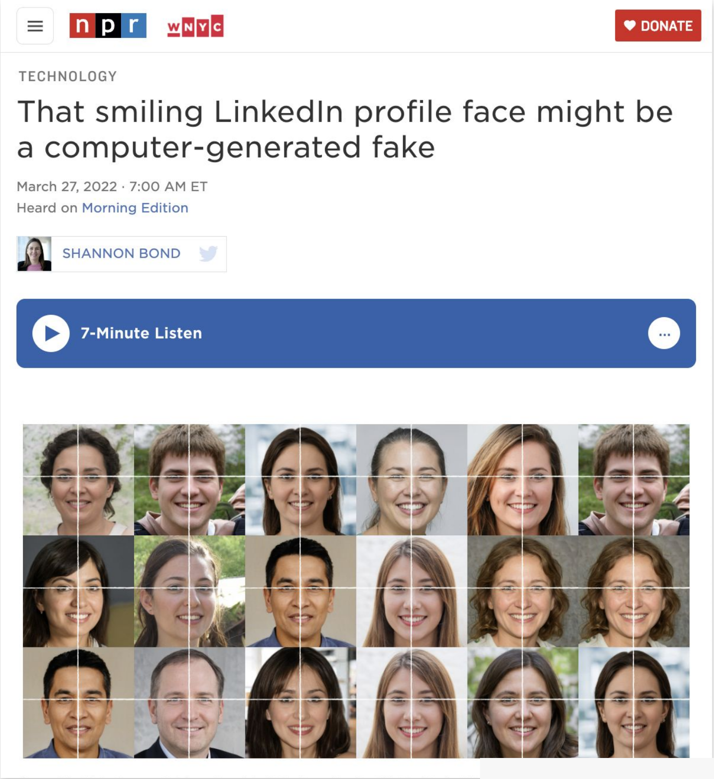
Sources:
Twitter 03/16/2022, ABC News, 06/24/2021, Evening Standard, 09/30/2022, MIT Technology Review 09/12/2021,

Entire face synthesis

→ Development of fully synthetic faces, which can be used to create fake identities and credentials

→ Lancaster University researcher: People can't tell the synthetic from real

(Source https://www.pnas.org/doi/10.1073/pnas.2120481119)



TECHNOLOGY

That smiling LinkedIn profile face might be a computer-generated fake

March 27, 2022 · 7:00 AM ET

Heard on Morning Edition

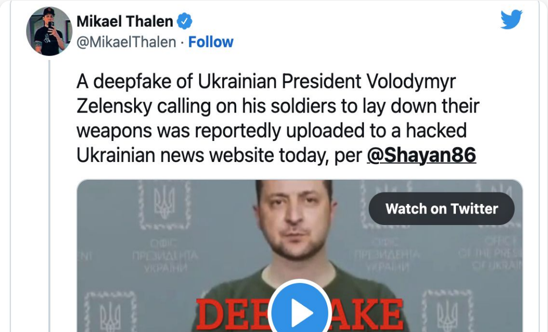
SHANNON BOND

Source:
NPR 03/27/2022

Identity/Expression swap

→ Replacing the face in a video or image with the face of another person

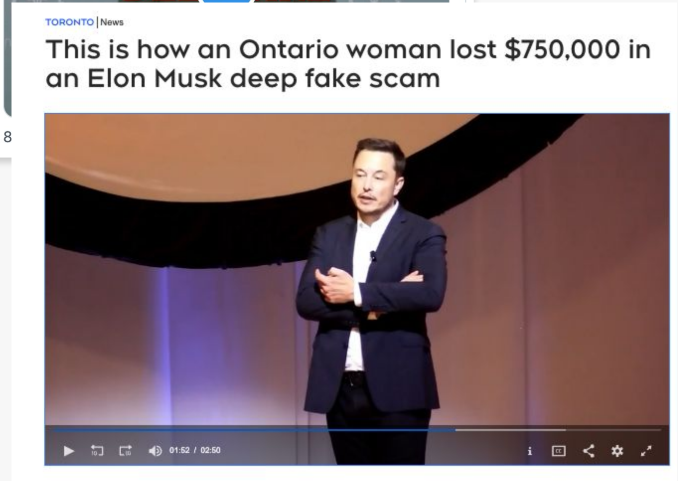
→ Applying the motion or expression from one person's face to an image or video of another person's face



Mikael Thalen
@MikaelThalen · Follow

A deepfake of Ukrainian President Volodymyr Zelensky calling on his soldiers to lay down their weapons was reportedly uploaded to a hacked Ukrainian news website today, per @Shayan86

Source:
Twitter 03/16/2022



TORONTO | News

This is how an Ontario woman lost $750,000 in an Elon Musk deep fake scam

Source:
CTV News 09/19/2022

[redacted]

Protections

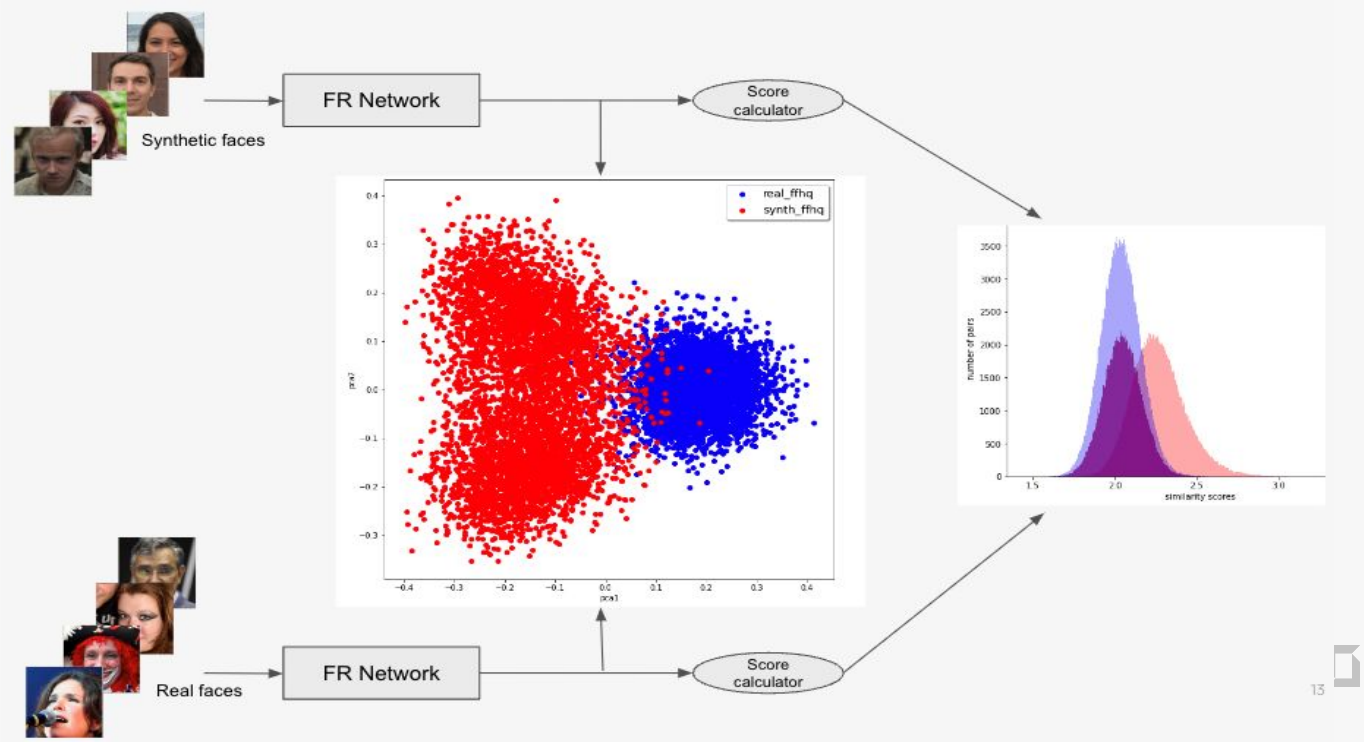
Protections: in the literature

A lot of research exists on:

→ Deepfake images detection/attribution and manipulation detection

→ Deepfake videos detection/attribution and manipulation detection

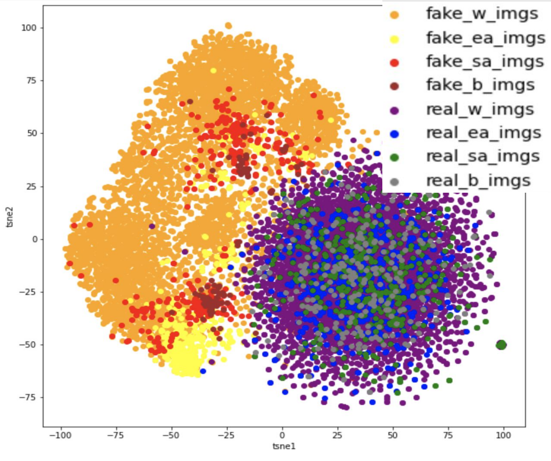
Our observations

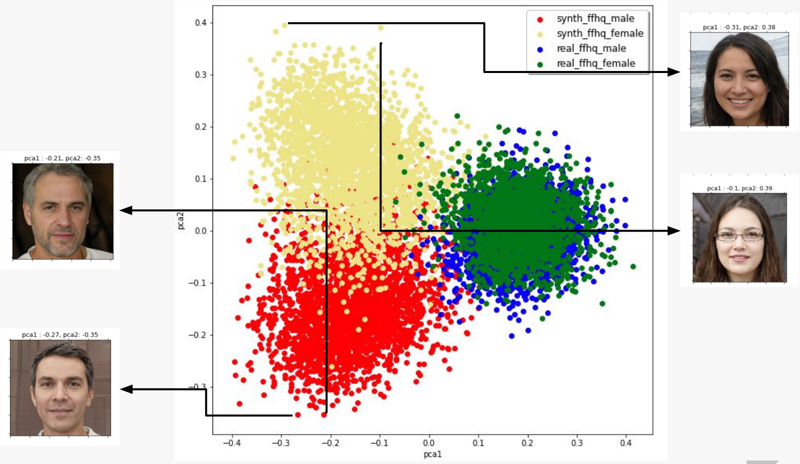


PARAVISION

Our observations

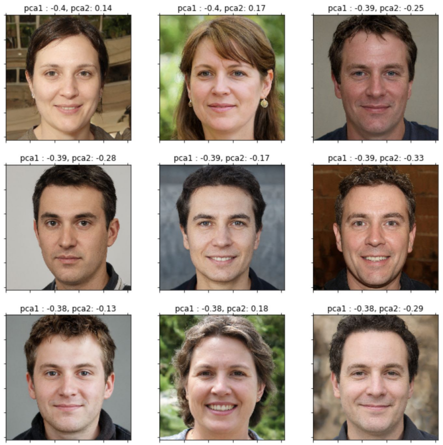
The synthetic embeddings seem more clusterable than the real images

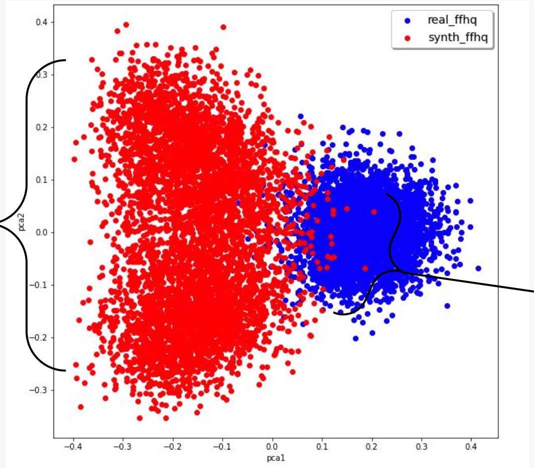




Our observations

Sampled average looking synthetic faces:


pca1 : -0.4, pca2: 0.14
pca1 : -0.4, pca2: 0.17
pca1 : -0.39, pca2: -0.25
pca1 : -0.39, pca2: -0.28
pca1 : -0.39, pca2: -0.17
pca1 : -0.39, pca2: -0.33
pca1 : -0.38, pca2: -0.13
pca1 : -0.38, pca2: 0.18
pca1 : -0.38, pca2: -0.29



Sampled real FFHQ faces:

Sampled sophisticated synthetic faces:

Paravision Synthetic Face Detection prototype

99.7% accuracy on detecting faces created with FFHQ and thispersondoesnotexist.com

Sample Detected Synthetic faces

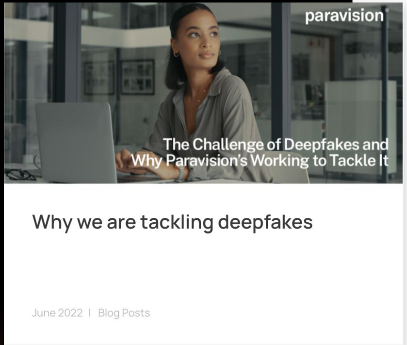
Sample Detected Real Faces

Progress in deepfakes generation and detection is being made

→ Paravision announced funding from a Five Eyes government for detecting deepfake videos

→ We are a building a multi-model production grade deepfake generator engine

→ We are building a production grade deepfake detector


paravision
The Challenge of Deepfakes and Why Paravision's Working to Tackle It

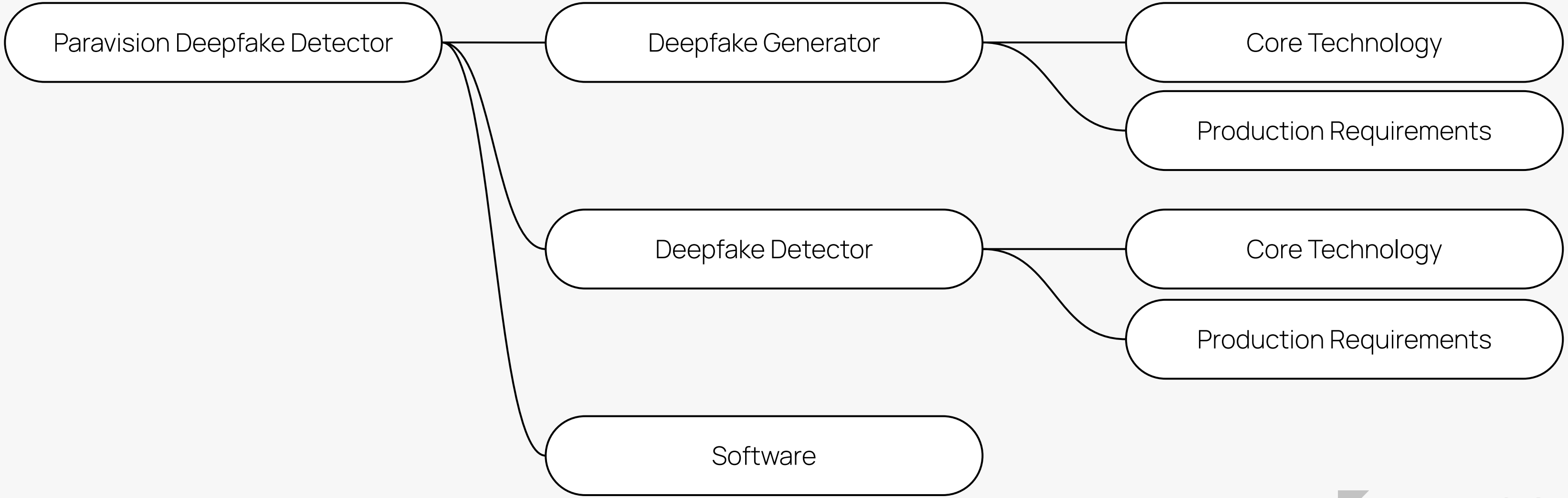
Why we are tackling deepfakes

June 2022 | Blog Posts

Paravision Selected for Government Deepfake Detection Program

June 2022 | Company News, Partnerships

Paravision Deepfake Detection Software

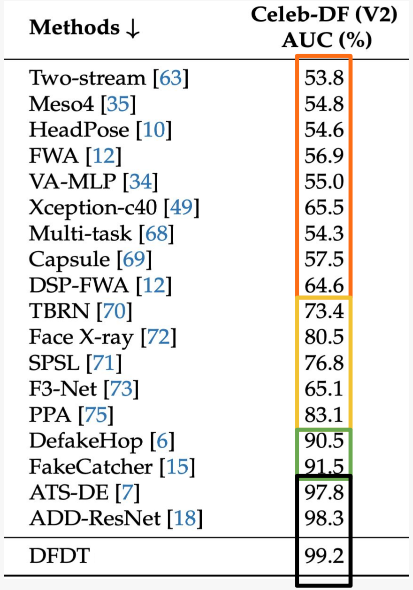


Deepfake detection in the literature

→ While detecting familiar deepfake videos is a simple binary classification problem

→ Detecting un-familiar deepfake videos is not so easy

→ Significant drop of accuracy can be seen when benchmarking on wild deepfake data, compared to the generalized deepfake dataset



Methods ↓	Celeb-DF (V2) AUC (%)
Two-stream [63]	53.8
Meso4 [35]	54.8
HeadPose [10]	54.6
FWA [12]	56.9
VA-MLP [34]	55.0
Xception-c40 [49]	65.5
Multi-task [68]	54.3
Capsule [69]	57.5
DSP-FWA [12]	64.6
TBRN [70]	73.4
Face X-ray [72]	80.5
SPSL [71]	76.8
F3-Net [73]	65.1
PPA [75]	83.1
DefakeHop [6]	90.5
FakeCatcher [15]	91.5
ATS-DE [7]	97.8
ADD-ResNet [18]	98.3
DFDT	99.2

mid 2017-mid 2021

mid 2021-Now

"DFDT: An End-to-End DeepFake Detection Framework Using Vision Transformer" (March 2022)

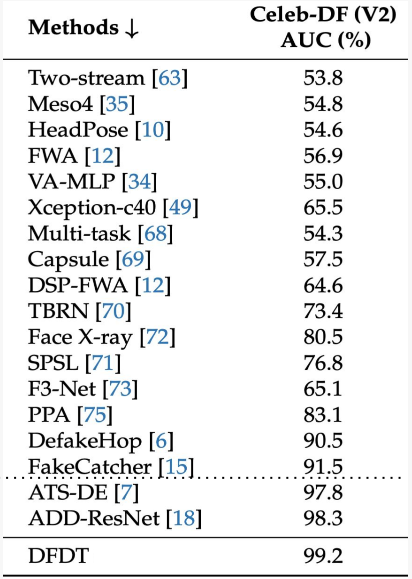
*celeb_df_v2 is a broadly used benchmark for generalized deepfake detection

Initial internal research and results

Model name	AUC for benchmarking sets		
	hq_FFPP	c40_FFPP	celeb_df_v2
dfd_paravision	98.8	92.5	96.1

trained on research datasets:

DeepFakeDetection + FaceForensics++



Methods ↓	Celeb-DF (V2) AUC (%)
Two-stream [63]	53.8
Meso4 [35]	54.8
HeadPose [10]	54.6
FWA [12]	56.9
VA-MLP [34]	55.0
Xception-c40 [49]	65.5
Multi-task [68]	54.3
Capsule [69]	57.5
DSP-FWA [12]	64.6
TBRN [70]	73.4
Face X-ray [72]	80.5
SPSL [71]	76.8
F3-Net [73]	65.1
PPA [75]	83.1
DefakeHop [6]	90.5
FakeCatcher [15]	91.5
ATS-DE [7]	97.8
ADD-ResNet [18]	98.3
DFDT	99.2

mid 2017-mid 2021

mid 2021-Now

"DFDT: An End-to-End DeepFake Detection Framework Using Vision Transformer" (March 2022)

Initial internal research and results

→In Paravision's **4 months** of R&D since mid 2022, our deepfake detector is performing at **96.1% accuracy** on generalized dataset

→The progress so far is fully based on **our internal models** instead of using off the shelf solutions

→We still need to test our detector against wild deepfakes

The main takeaways:

Synthetic faces

→ can be used to expand the training datasets

→ can be used to expand benchmarking the datasets, only if the FR model had seen enough synthetic faces of same source during training

FR networks pre-trained on real faces

→ can uncover several characteristics of synthetic faces, fully unsupervised

→ can make amazing deepfake detectors

PARAVISION

Thank you for your time.

Neda Eskandari, Ph.D.
Machine Learning Lead
neda@paravision.ai

Read more → paravision.ai

Dataset Attribution

Celeb DF V2
Li, Y. κ.ά. (2020) 'Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics', στο IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, United States.

FFHQ
A Style-Based Generator Architecture for Generative Adversarial Networks
Tero Karras (NVIDIA), Samuli Laine (NVIDIA), Timo Aila (NVIDIA)
https://arxiv.org/abs/1812.04948

FaceForensics++
Rössler, A. κ.ά. (2019) 'FaceForensics++: Learning to Detect Manipulated Facial Images', στο International Conference on Computer Vision (ICCV).

DeepFakes Detection
Dufour, N. κ.ά. (2019) 'DeepFakes Detection Dataset by Google & JigSaw'.