

# The Threat of Deepfakes & Their Impact on Organizations

**Dr. Yisroel Mirsky**

*Zuckerman Faculty Scholar*

*Tenure-track Lecturer*

*Head of the Offensive AI Research Lab*

*Department of Software and Information Systems Engineering, Ben-Gurion University*

*yisroel@bgu.ac.il*



**CBG**

Cyber@Ben-Gurion  
University of the Negev

**BIRD**

Israel-U.S.

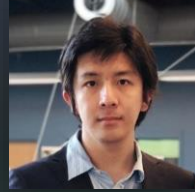
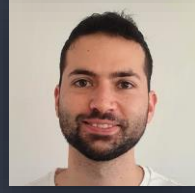
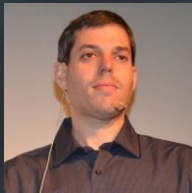
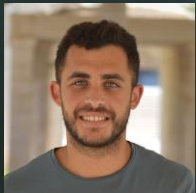
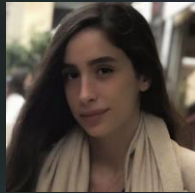
Binational Industrial Research  
and Development Foundation



<https://offensive-ai-lab.github.io/>



Principal Investigator



# Offensive AI Research Lab



<https://offensive-ai-lab.github.io/>

# Outline

- ▶ Introduction to **Deepfakes**
- ▶ Types of **Deepfakes**
  - ▶ Face & Voice
- ▶ The Threat Horizon

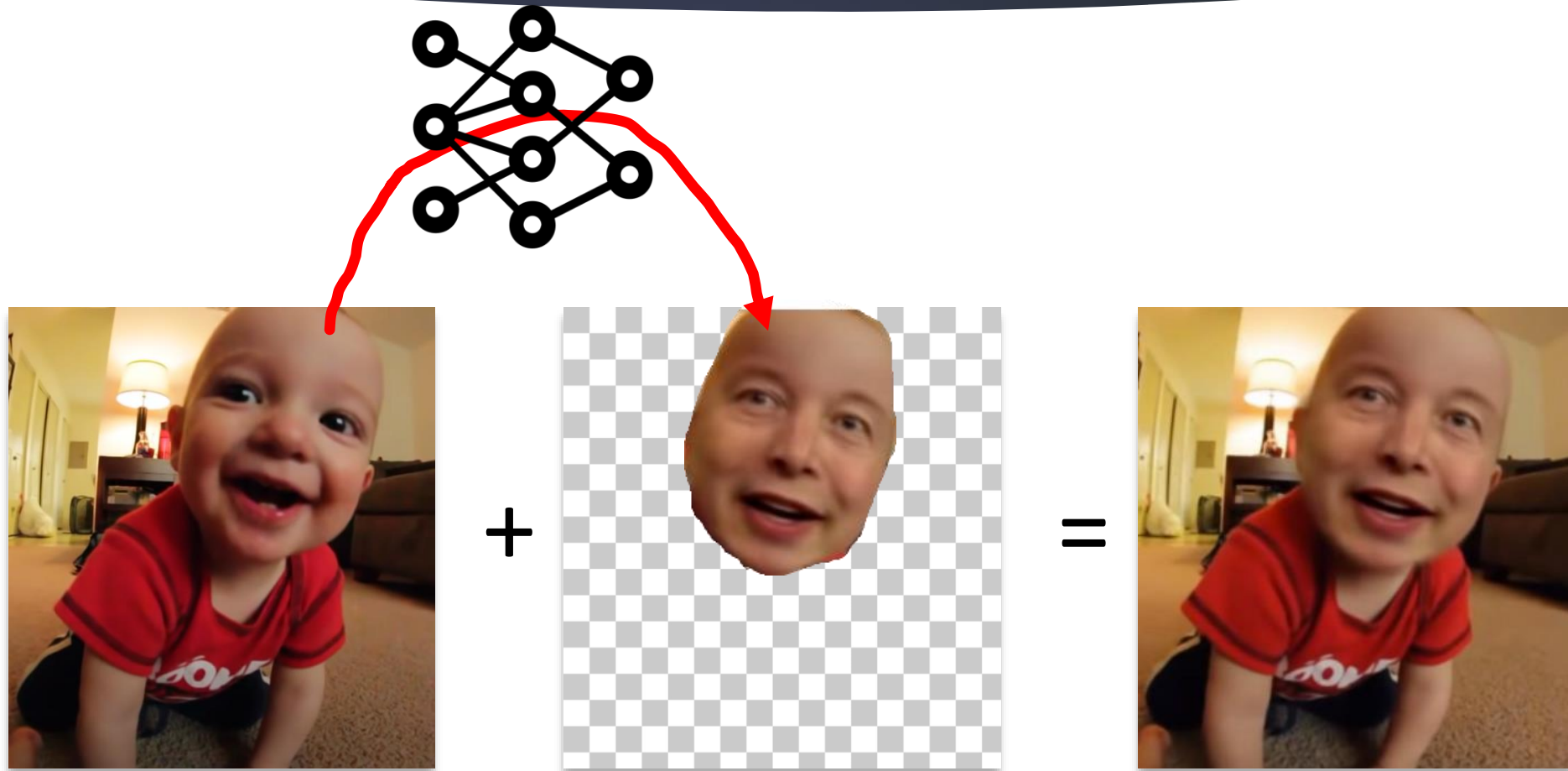




# Introduction



# The Common Deepfake



Can we trust our senses?



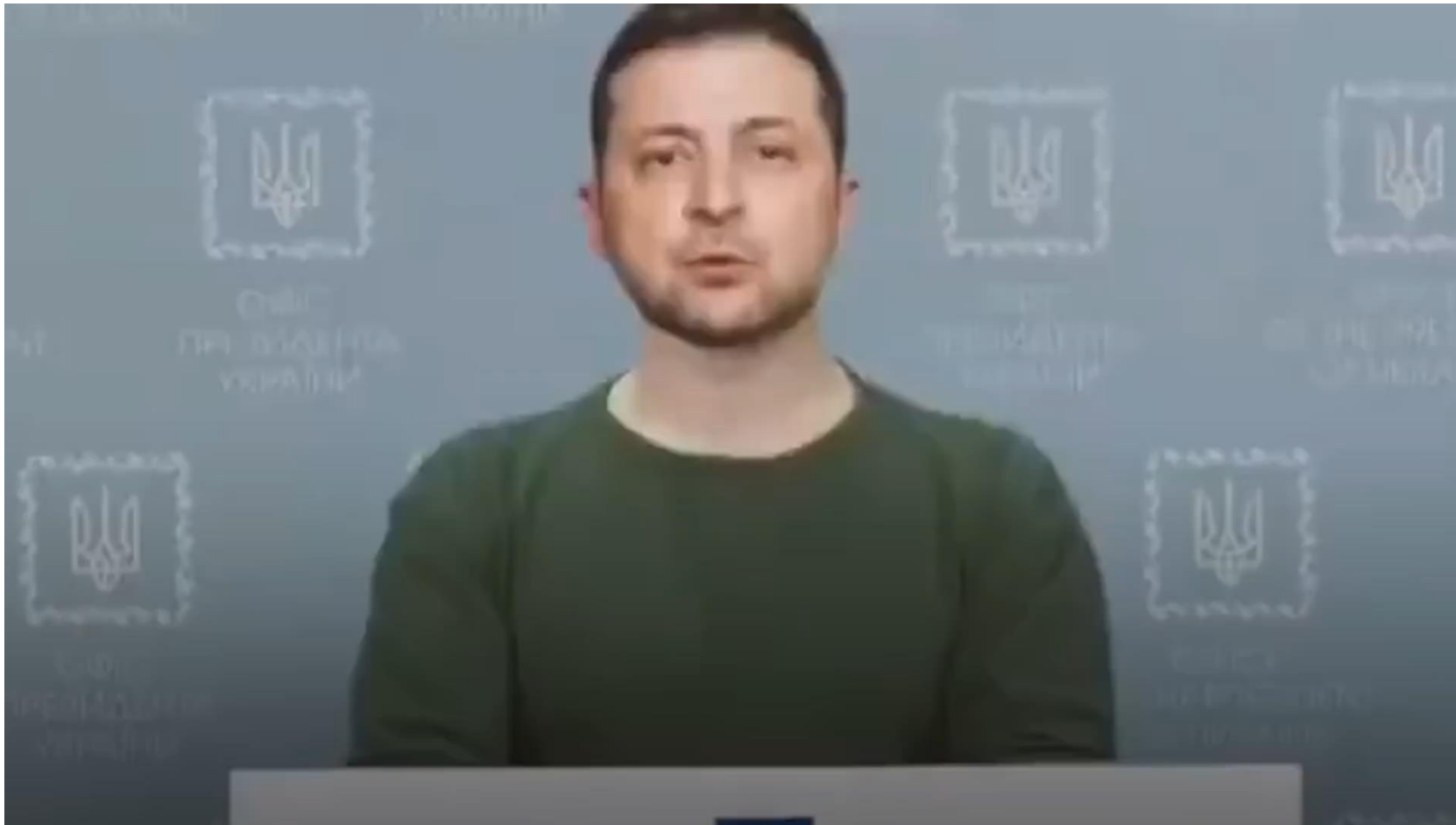
# Why should we Care?



We like our fun Nicholas Cage videos

# Why should we Care?

Impersonation and falsification  
of media is dangerous!



April 2022



# Why should we Care?

Impersonation and falsification of media is dangerous!



Social Engineering is the most common attack vector  
(humans are the often the weakest link!)

# Introduction

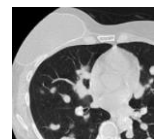
## What is a Deepfake?

Deep learning   Synthetic

“**Any** believable media generated by a deep neural network” [1]



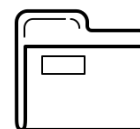
▶ **Video** (faces, surveillance, ...)



▶ **Images** (scenes, medical, ...)



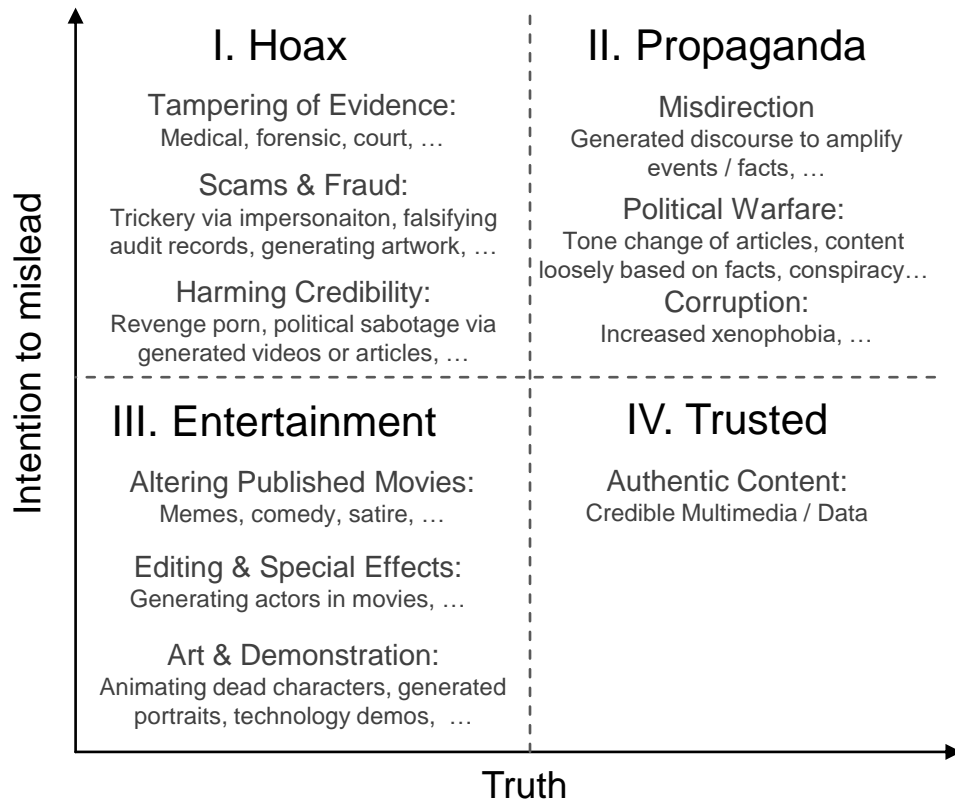
▶ **Audio** (voice, music, ...)



▶ **Records** (financial, logs, ...)

# Introduction

## Deepfake Information Trust Chart



<https://www.wired.com/story/elensky-deepfake-facebook-twitter-playbook/>

<https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/>

<https://www.forbes.com/sites/jess-edamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/?sh=665c7f9d2241>

**A Voice Deepfake Was Used To Scam A CEO Out Of \$243,000**

<https://www.engadget.com/netherlands-deepfake-video-chat-navalny-21266049.html>

<https://www.businessinsider.com/video-boris-johnson-endorses-jeremy-corbyn-in-convincing-deepfake-2019-11>

<https://www.mathejones.com/politics/2019/03/deepfake-gabon-ali-bongo/>

<https://www.theverge.com/2019/5/10/18540953/salvador-dali-lives-deepfake-museum>

<https://www.engadget.com/lucasli-m-hires-shamook-054904077.html>

<https://deepai.org/machine-learning-model/torch-srgan>

<https://shunsuke-saito.github.io/PiFuHD/>

# Introduction

## Malicious Use of “human” Deepfakes

### ▶ **Goals** - chronological:

- Old news*
- We are here*
- Emerging threats*
- ▶ Perform defamation
  - ▶ Blackmail, abuse (e.g., DF pornography)
  - ▶ Spread misinformation
  - ▶ Steal Money (Scams -abuse inherent trust)
  - ▶ Obtain Information (Social Engineering)
  - ▶ Cause an action (e.g., worker flip switch)
  - ▶ Tampering of evidence



### **Threat Actors:**

- ▶ state actor
- ▶ employee
- ▶ ...anybody



# Introduction

## When did it all start?

- ▶ **2014:** Ian Goodfellow develops GANs
- ▶ **2017:** Reddit user 'Deepfakes' swaps faces of women in explicit videos
- ▶ **2018:** BuzzFeed demonstrated misinformation risk with Obama video
- ▶ **2019 – Today:** Rapid development of deepfakes...



2014



2015



2016



2017



2018



... 2021



2022

# Deepfake Faces



# Deepfake Faces

## Deepfakes of Humans

$s$ : source identity,  $t$ : target identity

$x_s$ : source image,  $x_t$ : target image,  $x_g$ : generated image

### Four categories:

- ▶ Re-enactment
- ▶ Replacement
- ▶ Editing
- ▶ Synthesis

Source  $x_s$



Target  $x_t$



# Deepfake Faces

## Deepfakes of Humans

### Re-enactment



Generated  $x_g$



●: Always  
○: Sometimes

**Transfers:**  
Gaze  
(dubbing) Mouth  
Expression  
(face or body) Pose  
Identity

	Gaze	Mouth	Expression	Pose	Complete
Gaze	●		○		○
(dubbing) Mouth		●	○		●
Expression		○	●		●
(face or body) Pose				●	●
Identity					●




# Deepfake Faces

## Deepfakes of Humans


### Replacement

Source  $x_s$       Target  $x_t$       Generated  $x_g$



●: Always  
○: Sometimes

Drives:	Transfer	Swap
Gaze		○
(dubbing) Mouth		●
Expression		●
(face or body) Pose		●
Identity	●	●

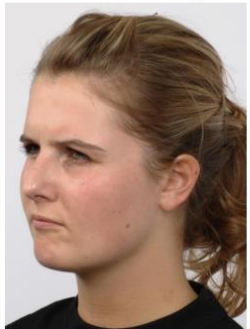


# Deepfake Faces

## Deepfakes of Humans

### Editing

Source  $x_s$



Target  $x_t$



Generated  $x_g$



*Hair*

*Article*

*Age*

*Beauty*

*Ethnicity*

# Deepfake Faces

## Deepfakes of Humans

Synthesis



[Thispersondoesnotexist.com](http://Thispersondoesnotexist.com)

We will focus on **Re-enactment** and **Replacement**

# Facial Re-enactment & Replacement

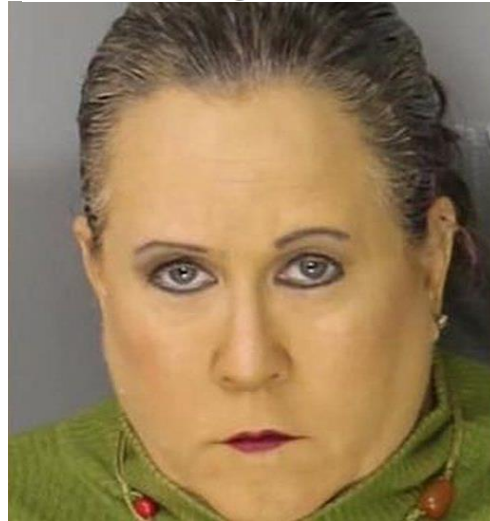
**A real threat...**

## European MPs targeted by deepfake video calls imitating Russian opposition

Politicians from the UK, Latvia, Estonia and Lithuania tricked by fake meetings with opposition figures



## Mother 'used deepfake to frame cheerleading rivals' **BBC**



# iNews

## Valentines Day: Online romance fraud nearing an 'industrial scale' as criminals embrace deepfake technology

EXCLUSIVE

Nearly 9,000 victims of dating scams reported being targeted to police last year and lost a total of 97.2m, figures show



Romance fraud is nearing an 'industrial scale', experts have warned (Photo: Alamy)

## Hackers reportedly deepfaked a Binance exec to carry out listing scams

The scammers seemingly used Patrick Hillmann's media appearances to ape the chief comms officer's image.



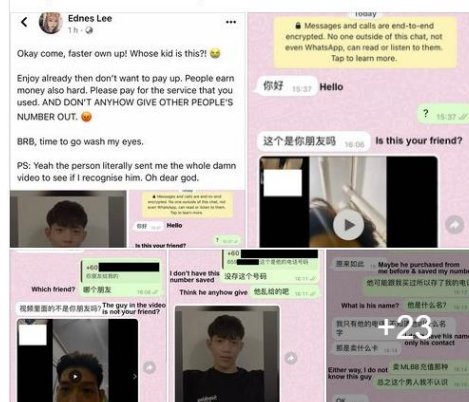
K. Holt  
@krisholt  
August 23, 2022  
11:47 AM



In this article: binance, news, near deepfake



## Singaporean man's face ends up in deepfake porn after he refuses to pay hacker \$5,800 **yahoo/news**



10 March 2021

PIN Number  
210310-001

Please contact the FBI with any questions related to this Private Industry Notification at either your local Field Office.

Local Field Offices:  
[www.fbi.gov/contact-us/field-offices](http://www.fbi.gov/contact-us/field-offices)

The following information is being provided by the FBI, with no guarantees or warranties, for potential use at the sole discretion of recipients to protect against cyber threats. This data is provided to help cyber security professionals and system administrators guard against the persistent malicious actions of cyber actors. This PIN was coordinated with DHS-CISA.

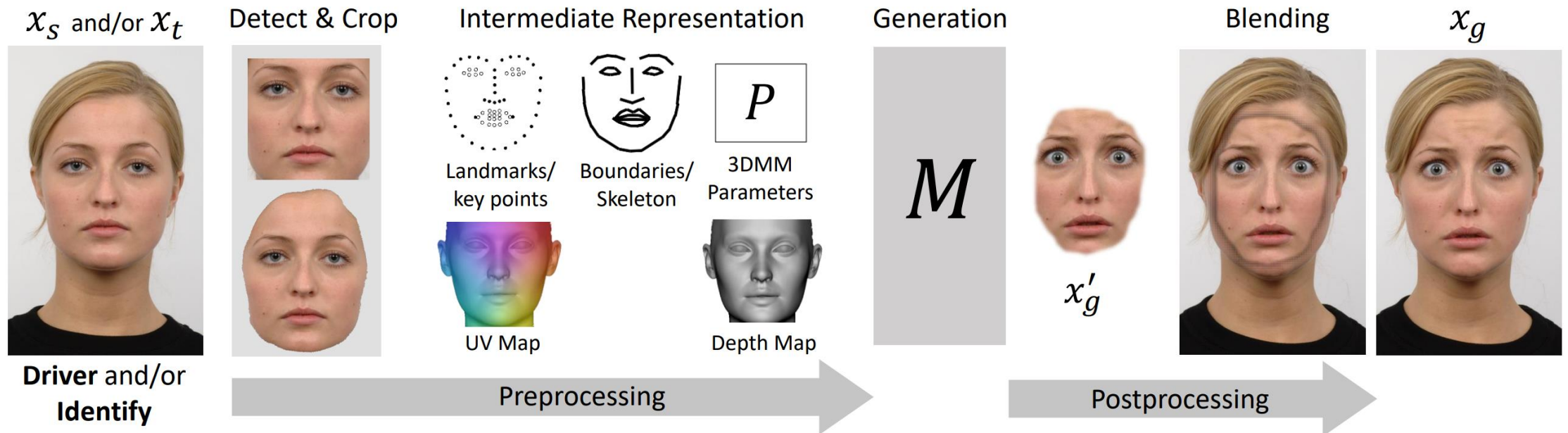
This PIN has been released **TLP:WHITE**. Subject to standard copyright rules, **TLP:WHITE** information may be distributed without restriction.

**Malicious Actors Almost Certainly Will Leverage Synthetic Content for Cyber and Foreign Influence Operations**



# Deepfake Faces

## DF Creation Basics: The pipeline

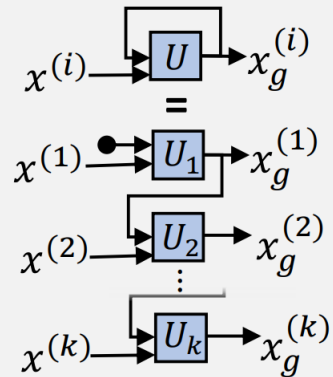


Pipeline varies depending on the scope of  $x'_g$  (whole scene, head, or just face)

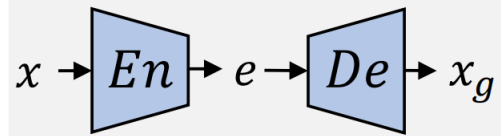
# Deepfake Faces

## Basic Architectures

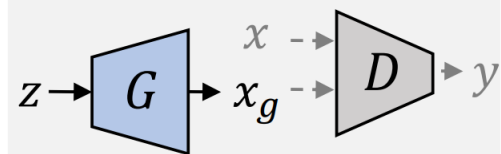
### RNN



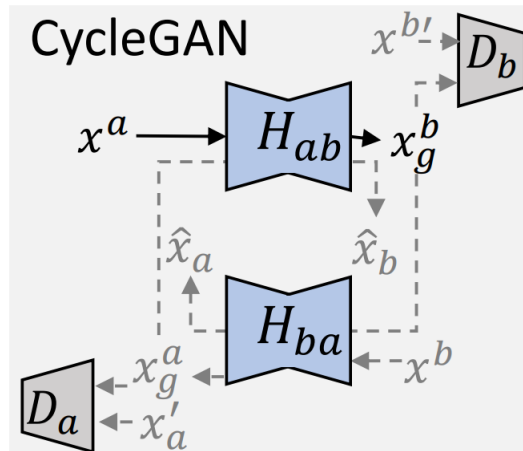
### Encoder Decoder



### Vanilla GAN



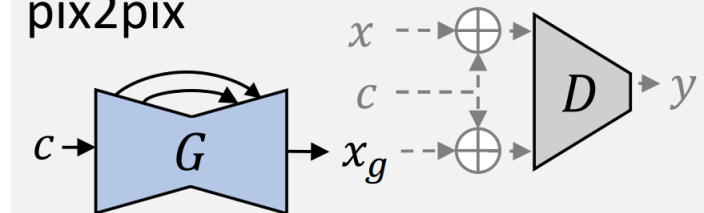
### CycleGAN



### Networks

- Generative
- Discriminator

### pix2pix

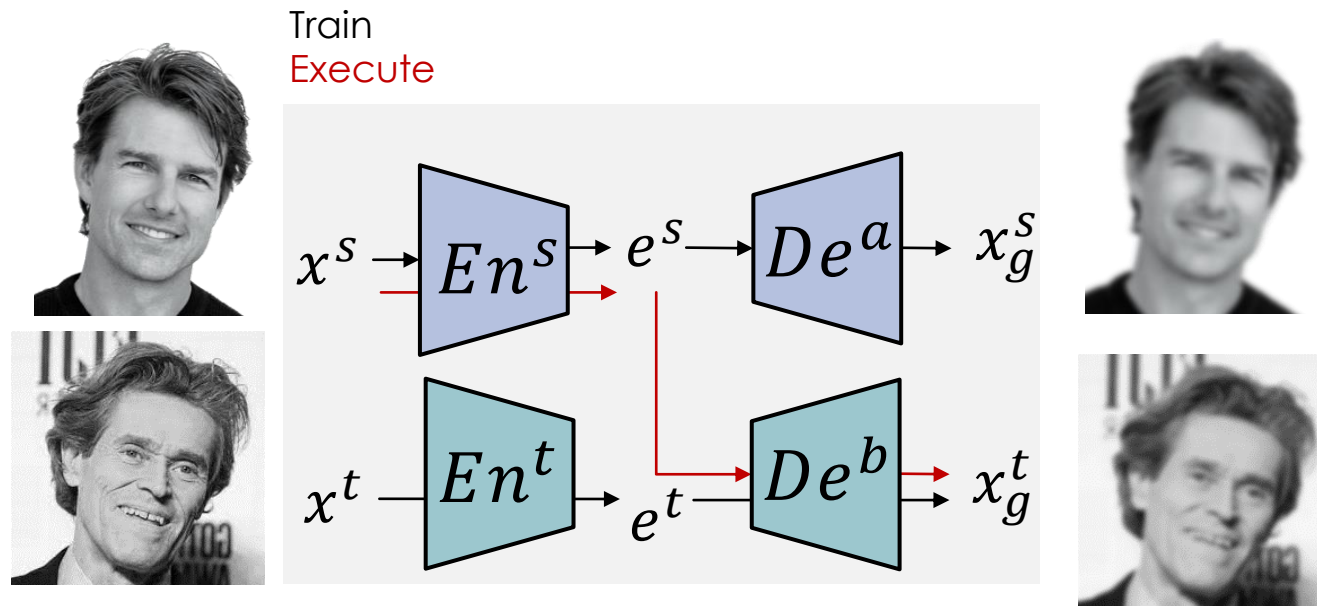


# Deepfake Faces

## Six Ways to Drive a Face (DF Design Patterns)

1. Let  $M$  learn the mapping itself from a direct representation

### Example Model:



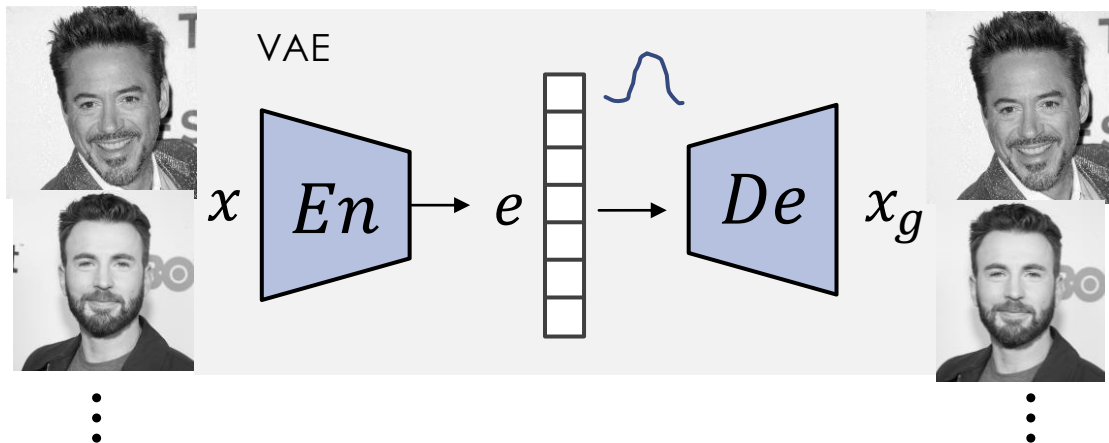
# Deepfake Faces

## Six Ways to Drive a Face (DF Design Patterns)

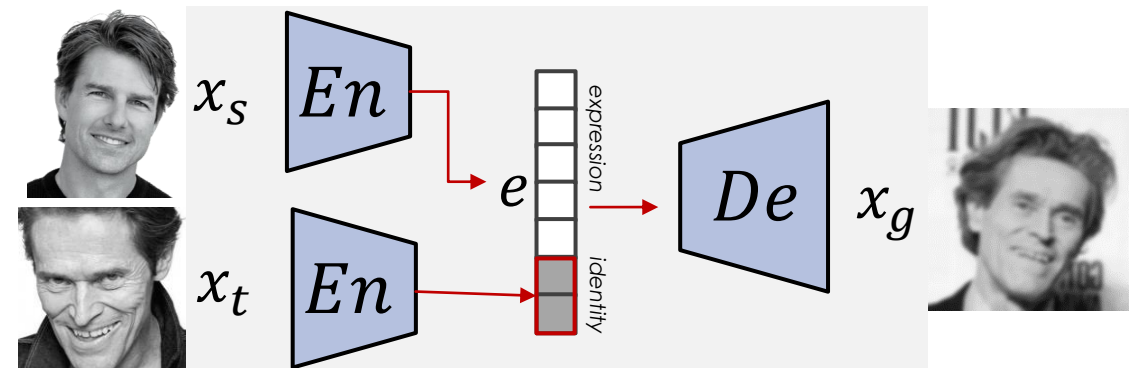
2. Train  $En$  to disentangle identity from expression, then modify/swap encoding before  $De$

### Example Model:

Train



Execute





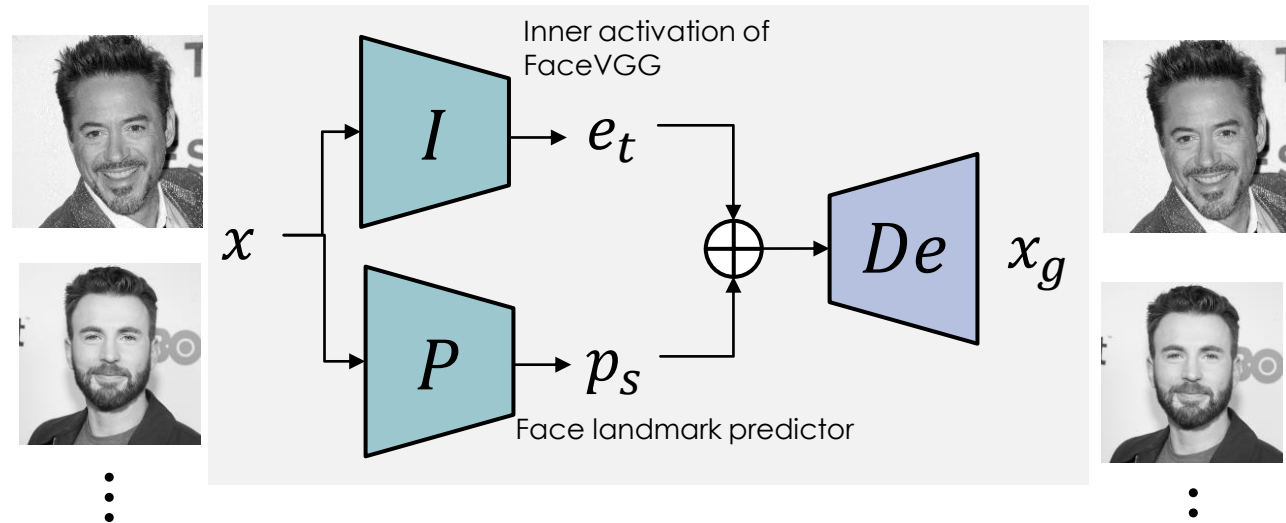
# Deepfake Faces

## Six Ways to Drive a Face (DF Design Patterns)

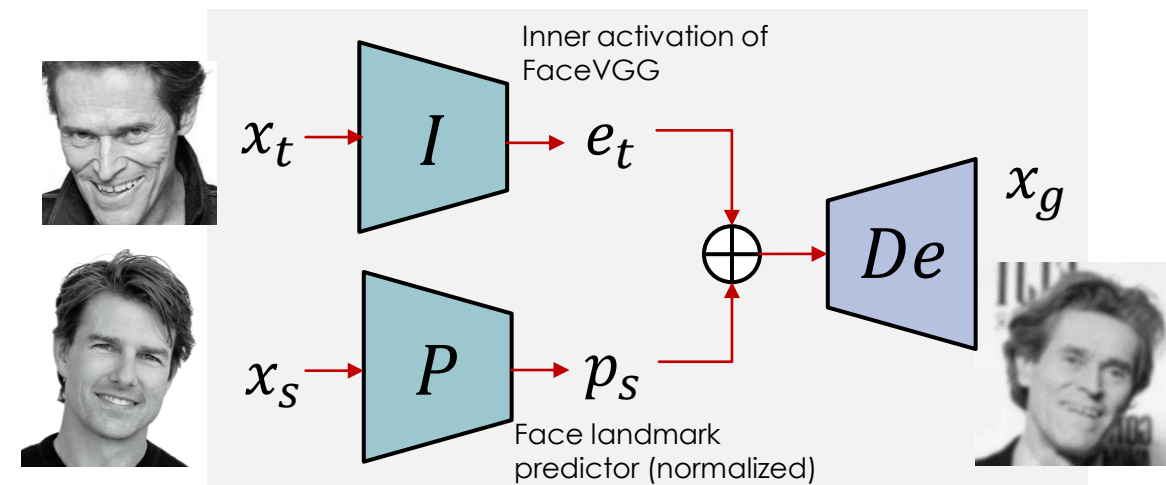
3. Add additional encoding (e.g., AU, LANDMARK, embedding) before  $De$

### Example Model:

Train



Execute

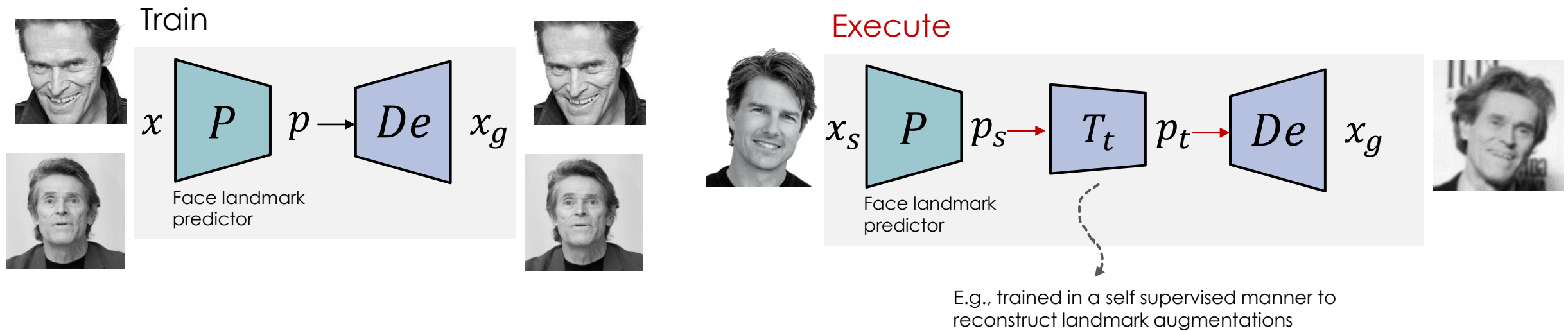


# Deepfake Faces

## Six Ways to Drive a Face (DF Design Patterns)

- Convert intermediate representation to that of  $t$  before  $G$

### Example Model:

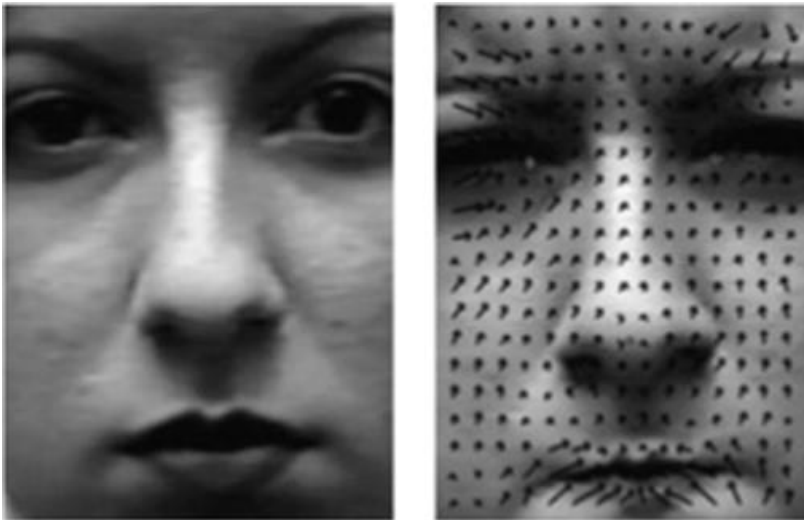


# Deepfake Faces

## Six Ways to Drive a Face (DF Design Patterns)

- Use optical flow (from previous/other frame) for video

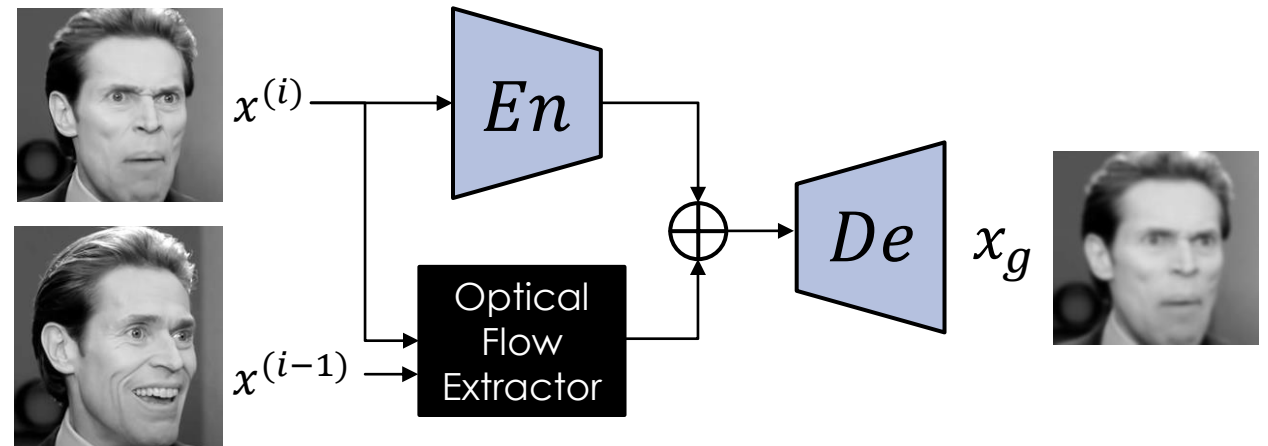
**Optical Flow** measures the pixel-wise displacement between two frames



E.g., library OpenCV

## Example Model:

Train (ED approach)

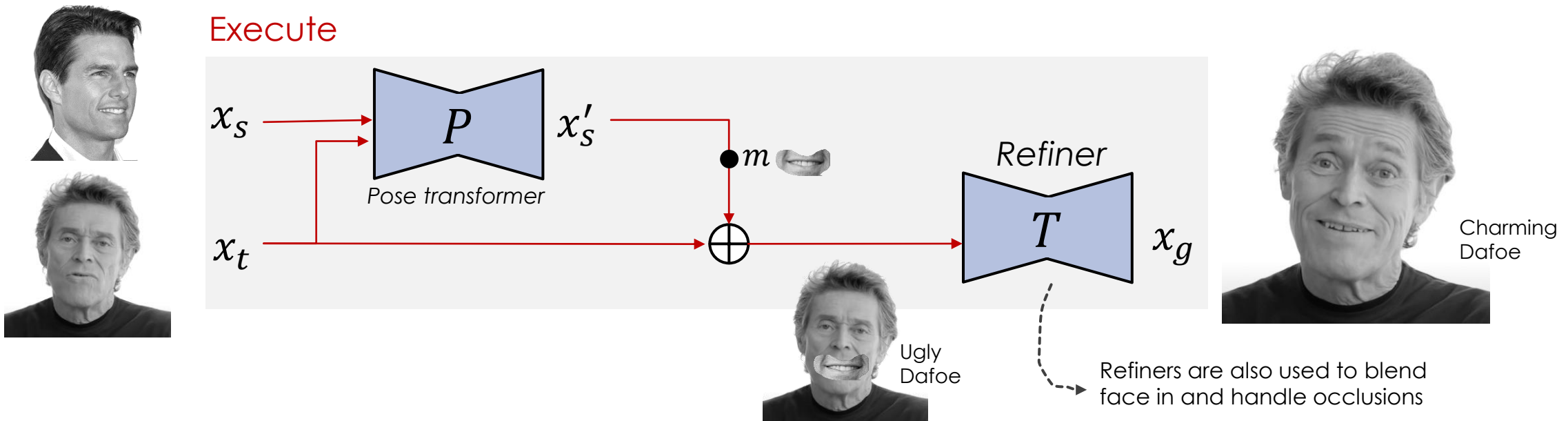


# Deepfake Faces

## Six Ways to Drive a Face (DF Design Patterns)

6. Create composite input from several representations, then refine concatenation with another network (e.g., pix2pix)

### Example Model: Dubbing

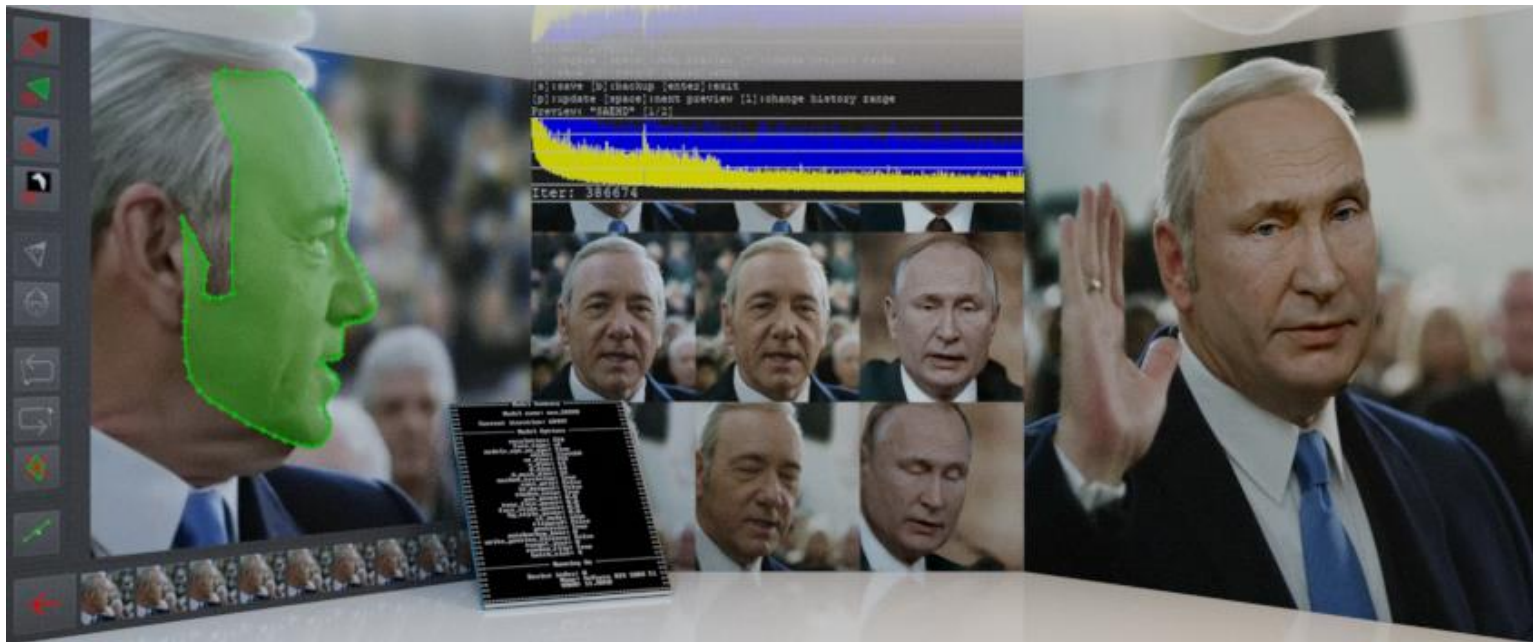


# Deepfake Faces

## Example for Replacement/face swap

The Most Popular face swap tool:

### Deep Face Lab



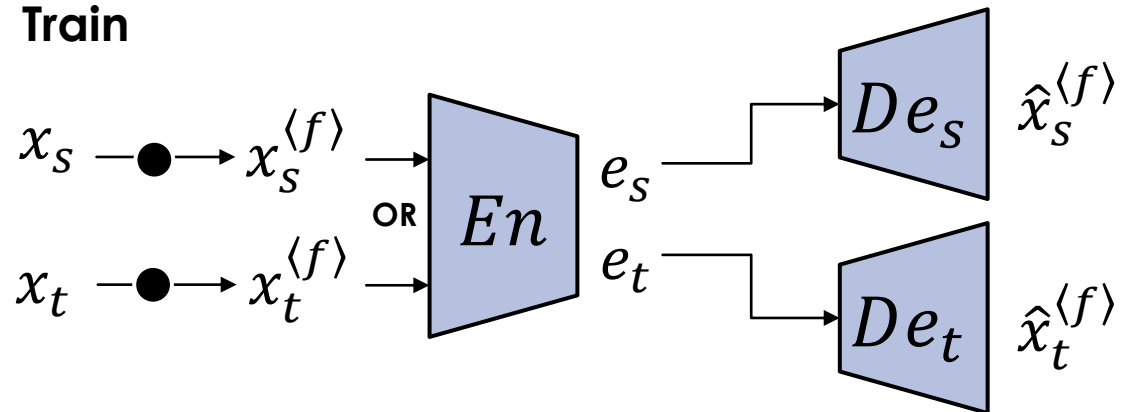


# Deepfake Faces

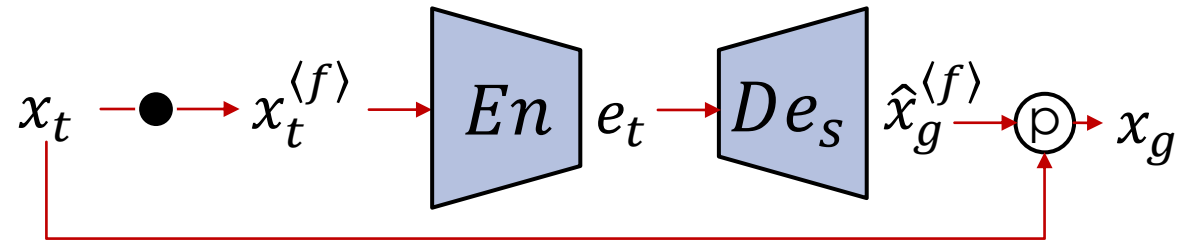
## The Most Popular Face Replacement (one-to-one)

- ▶ Used by first Reddit deepfake, still used in tools like DeepFaceLab
  - ▶ Shared  $En$  means no image pairing in training!

### Train



### Execute



# Deepfake Faces

Mouth Re-enactment (dubbing)

## Attack Goals

### Misinformation



“Opposition leader bob is right!”



“I hate <racist comment>”



“Today TESLA stocks fell 10%”

### Social Engineering



“Hey Joe, one which server do we keep the credentials?”



“Turn off the firewall for 10 minutes, I’m doing some tests”

# Deepfake Faces

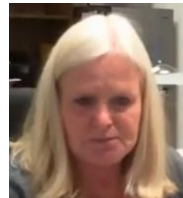
## Mouth Re-enactment (dubbing)

### The Pipeline

Target

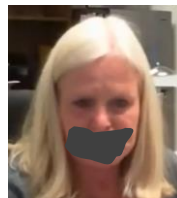


Extraction



$x_t$

Pre-processing



$x_t^{(m)}$



$x_t^m$  optional reference

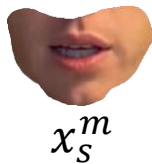
Generation



Post-processing



Source  
(driver)



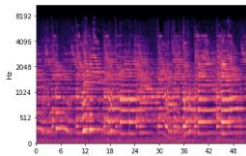
$x_s^m$



$x_s^a$



$l_s, l_t$



$a_s$

# Deepfake Faces

Mouth Re-enactment (dubbing)

**Examples** (they pass the Turing Test!)



Vougioukas K., et al. Realistic Speech-Driven Facial Animation with GANs, 2019



# Deepfake Voices

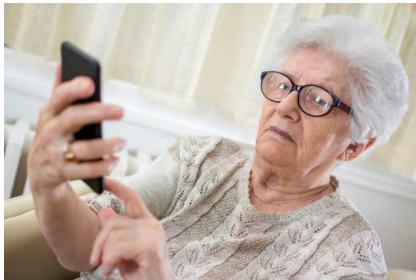


© DW

# Deepfake Voices

## Attack Goals

Scams



“Mom, I need help!”



“Transfer that 100k right away!”

Social Engineering



“The levels are too low...”



“What’s the IP of our portal?”

Authentication



“Alexa, unlock front door”  
“I’m Robert, send my new SIM to...”

# Deepfake Voices

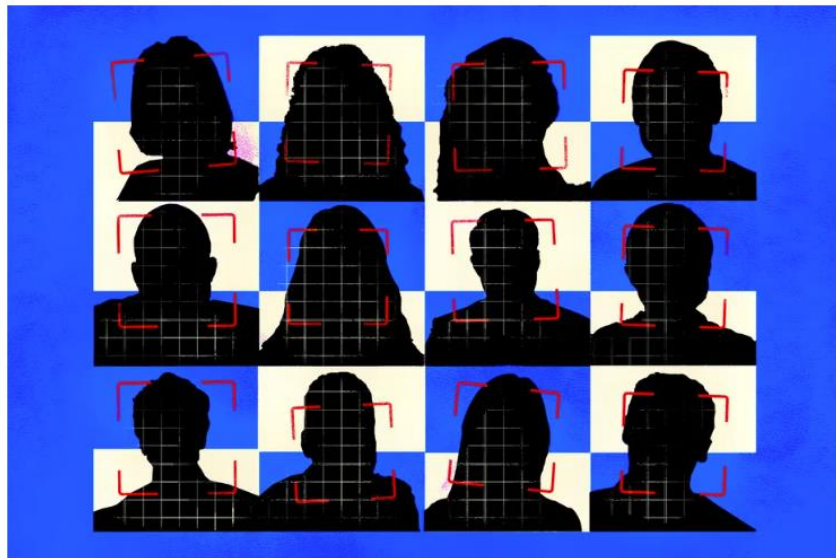
## THE VERGE

### Liveness tests used by banks to verify ID are 'extremely vulnerable' to deepfake attacks

Attackers can simply swap their face for another

By James Vincent | May 18, 2022, 9:00am EDT

f t s SHARE



## Forbes

INNOVATION

### Beware The Tactics Used For CEO Fraud By BEC Scammers

Stu Sjouerman Forbes Councils Member  
Forbes Technology Council COUNCIL POST | Membership (Fee-Based)

May 18, 2022, 08:15am EDT

f Stu Sjouerman is the founder and CEO of KnowBe4 Inc., a security awareness training and simulated phishing platform.



GETTY

Business email compromise (a.k.a. CEO fraud) is the highest-grossing type of cybercrime, according to the FBI's IC3 Internet Crime Report 2021. More than a third of all cybercrime losses can be attributed to BEC scams, causing about \$2.4 billion in losses to U.S. businesses last year, a 33% increase from 2020 and a tenfold increase from just seven years ago. Between 2013 and 2019, CEO fraud reportedly cost the economy a

## Forbes

CONSUMER TECH

### A Voice Deepfake Was Used To Scam A CEO Out Of \$243,000

Jesse Damiani Contributor @  
Iran Postreality Labs, a new media art advisory & curatorial studio.

Sep 3, 2019, 04:42pm EDT

Listen to article 3 minutes

This article is more than 2 years old.



Anonymous hacker programmer uses a laptop to hack the system in the dark. Creation and infection of ... [1] GETTY

It's the first noted instance of an artificial intelligence-generated voice deepfake used in a scam.

2019

## Forbes

### Fraudsters Cloned Company Director's Voice In \$35 Million Bank Heist, Police Find

Thomas Bruester Forbes Staff  
Associate editor at Forbes, covering cybercrime, privacy, security and surveillance.

Oct 14, 2021, 07:02am EDT



AI voice cloning is used in a huge heist being investigated by Dubai investigators, amidst warnings about cybercriminal use of the new technology.

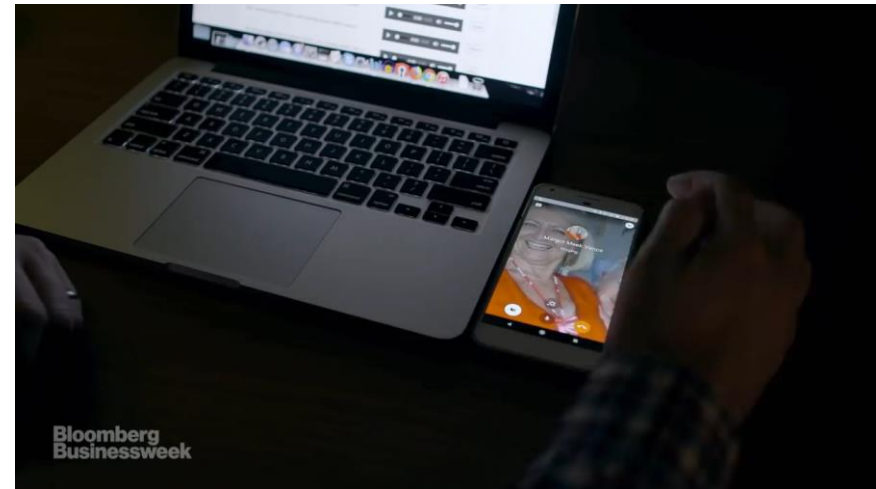
2021

# Deepfake Voices

## Voice Cloning via TTS Services



Requires: 10 minutes of audio





# Deepfake Voices

## Voice Cloning via TTS

### Services

Are they used by attackers? Almost certainly

Many require accepting terms of use

“use only voices you have right to” 🙄



Others will only train on a reading script

But... attacker could collect words from past recordings

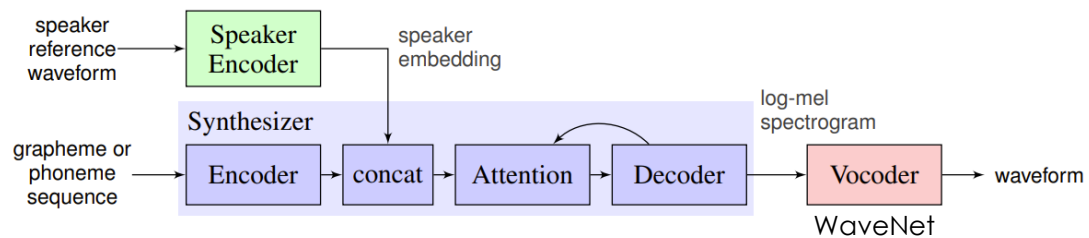




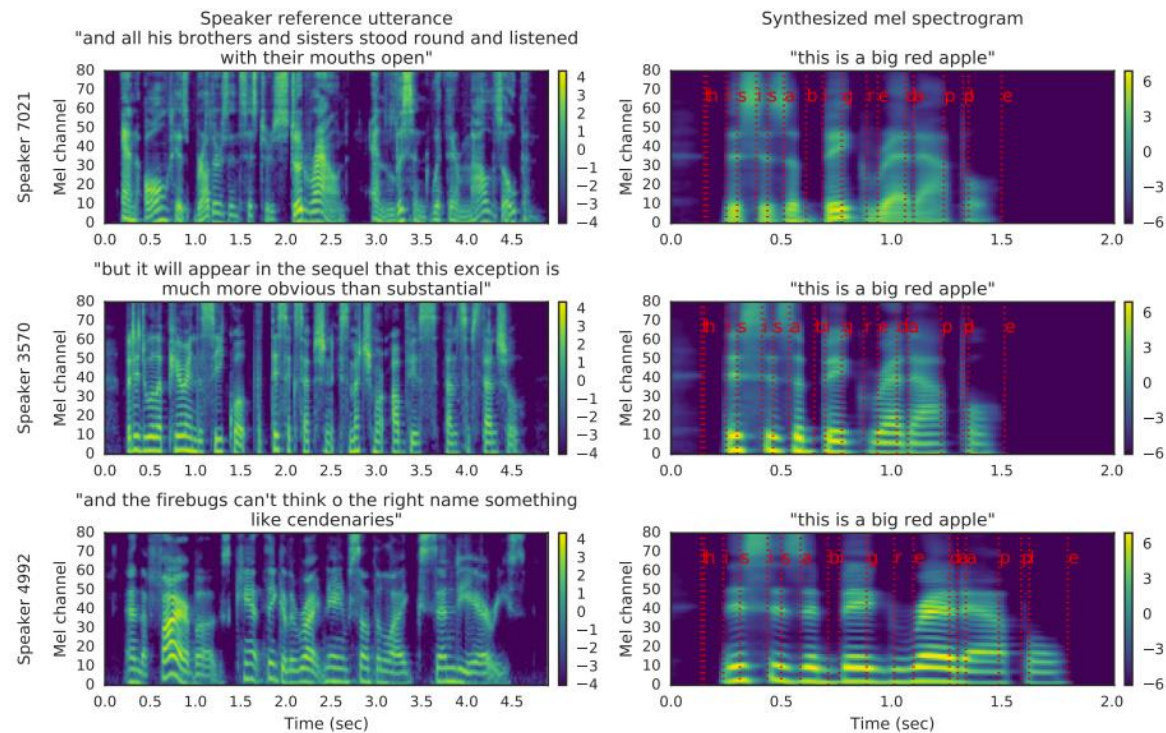
# Deepfake Voices

## Voice Cloning via TTS

### Zero-shot (only 3 seconds!)



Attention is used to help network align sequence to audio

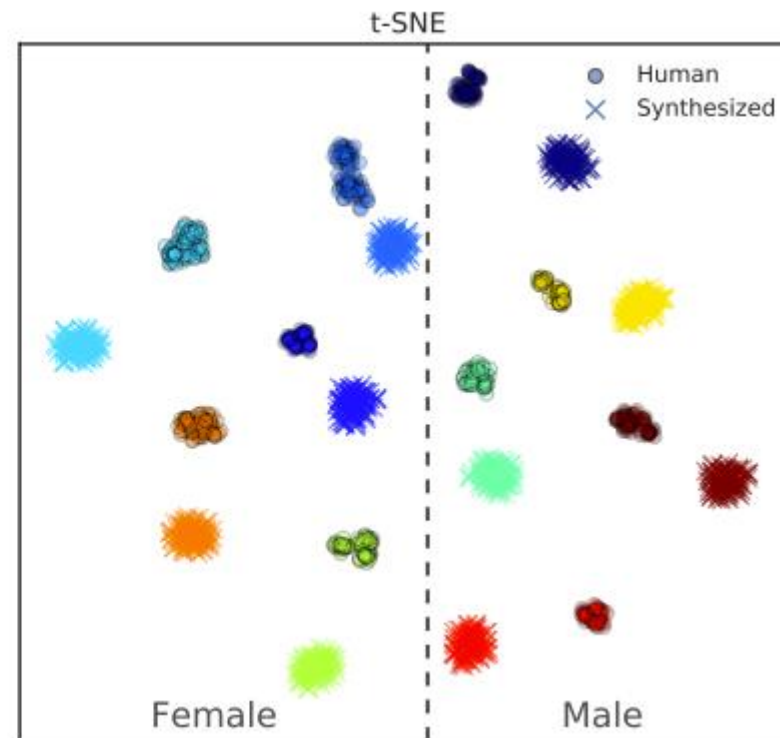
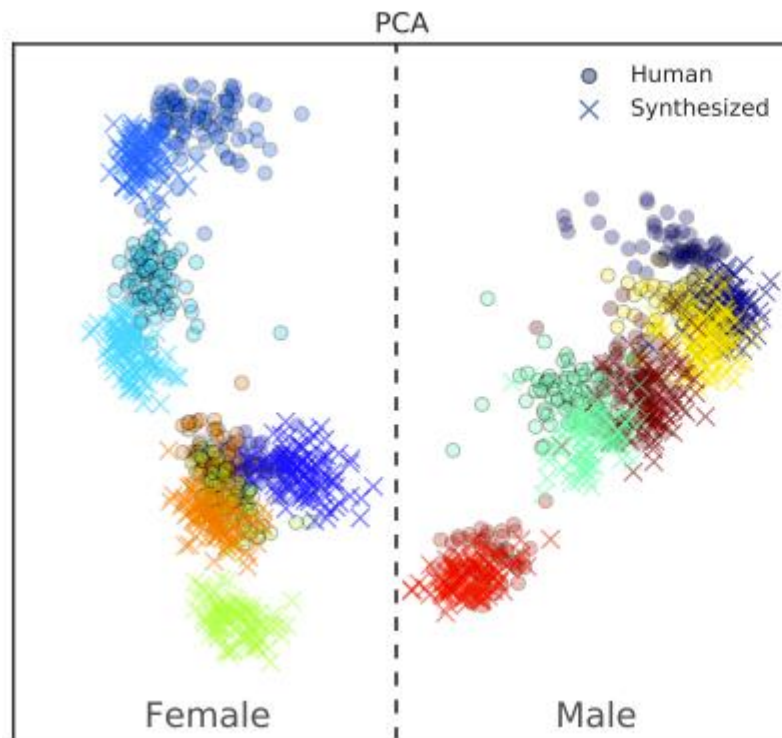


# Deepfake Voices

## Voice Cloning via TTS

### Zero-shot (only 3 seconds!)

Fake and Real identities fall close in embedding space



# Deepfake Voice

## Voice Cloning via TTS

Remember those money transfer scams?



CONSUMER TECH

## A Voice Deepfake Was Used To Scam A CEO Out Of \$243,000

Jesse Damiani Contributor   
I run Postreality Labs, a new media art advisory & curatorial studio.

Follow

Sep 3, 2019, 04:42pm EDT

 Listen to article 3 minutes 

 This article is more than 2 years old.



Anonymous hacker programmer uses a laptop to hack the system in the dark. Creation and infection of ... [+] GETTY

**It's the first noted instance of an artificial intelligence-generated voice deepfake used in a scam.**

# Deepfakes in **Media Editing**





# Deepfakes in Media Editing

## DALL-E 2, Stable Diffusion, ...

- ▶ Can modify existing images too

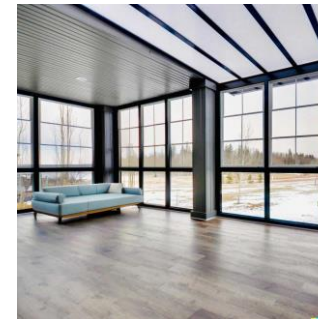
Original



(1)



(2)

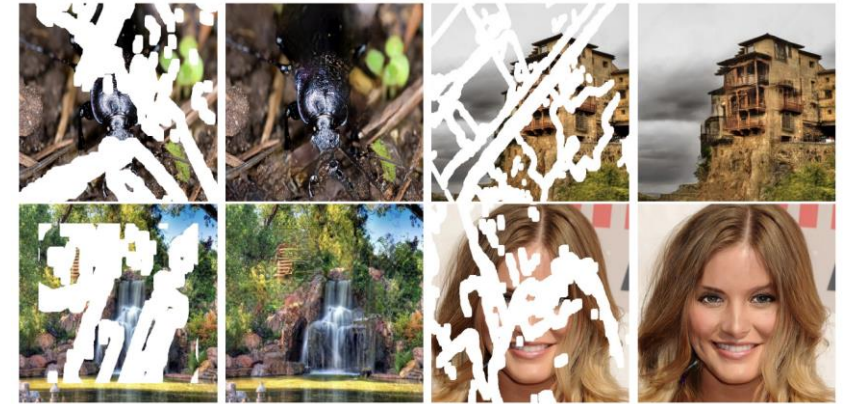
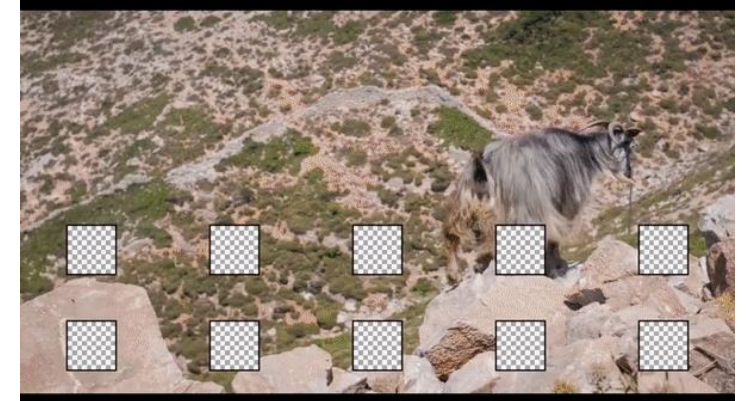




# Deepfakes in Media Editing

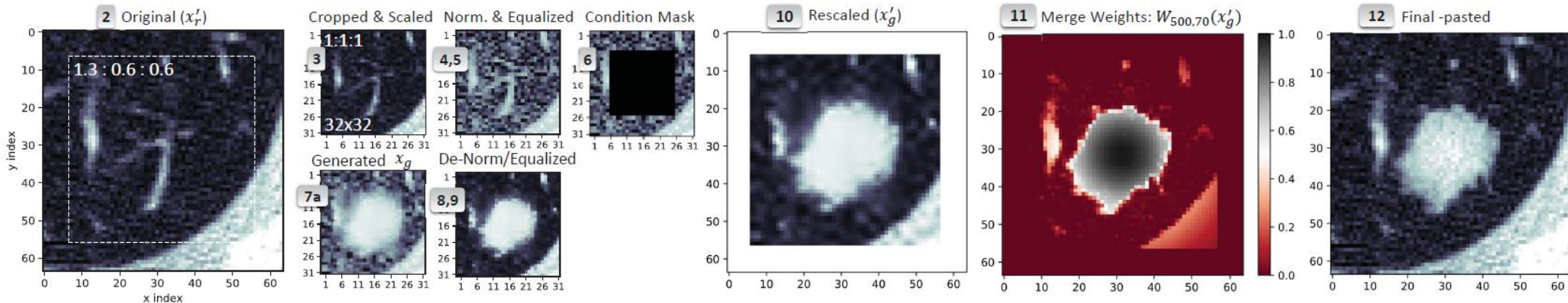
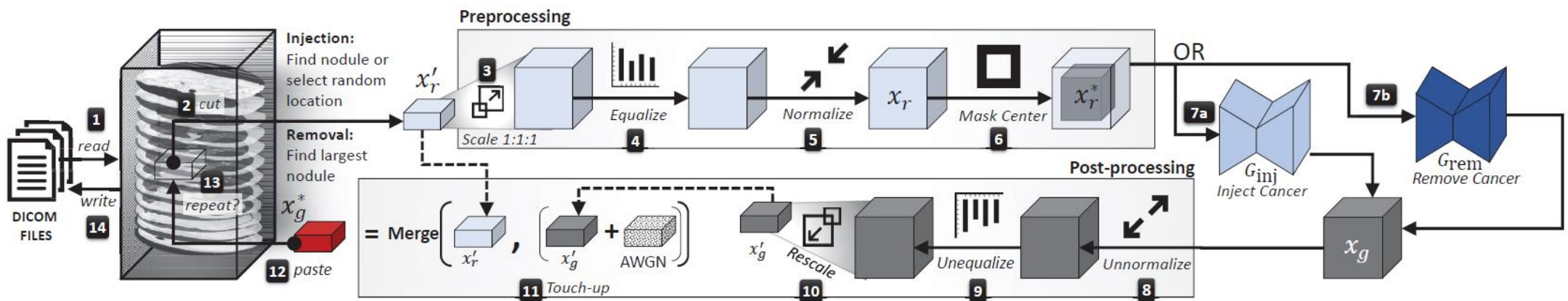
## Inpainting

**Definition:** The task of filling in missing content



# Deepfakes in Media Editing

## CT-GAN





# Example: Injecting and Removing Lung Cancer with Deep Learning



# How can an Attacker Accomplish This?

(one possible way)

- 1) Build a man-in-the-middle device
- 2) At night, plant the device near the scanner
- 3) Intercept CT scans, and manipulate them

*Demonstration in real Hospital...*

# Deepfake Defences





# Deepfake Defences

## Popular Techniques

### Undirected Approaches

- ▶ Classification
- ▶ Anomaly Detection



Evaded with  
adversarial noise

**ML given all features**  
(learns own features)

### Directed Approaches (Artifact-Specific)



**ML focused on specific features**

# Deepfake Defences

## 2. Directed Approaches *(using ML on specific artifacts)*

### Seven types of Artifacts:

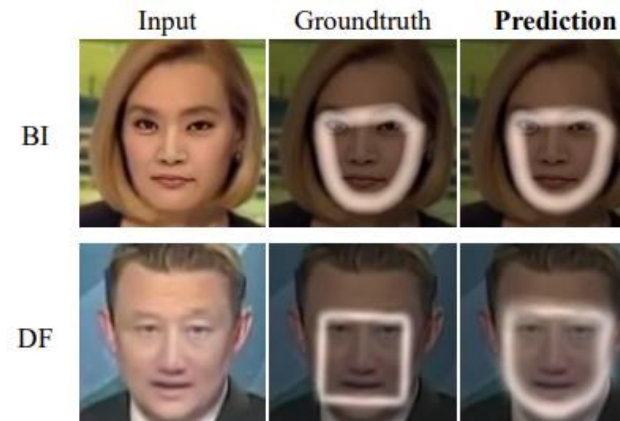
#### ► Spatial

1. Blending
2. Environment
3. Forensics

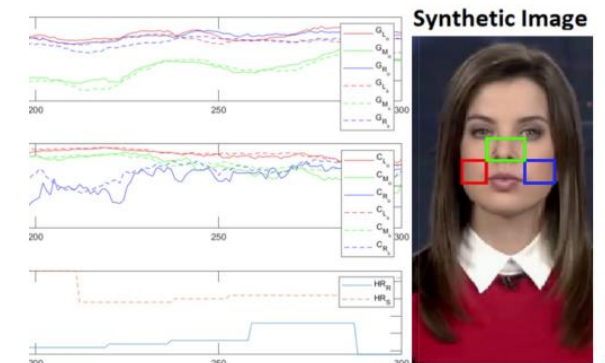
Concepts overlap with classical Forensics (edge+region)

#### ► Temporal

4. Behaviour
5. Physiology
6. Synchronization
7. Coherence



Li L. Et al. Face X-ray for More General Face Forgery Detection. 2020



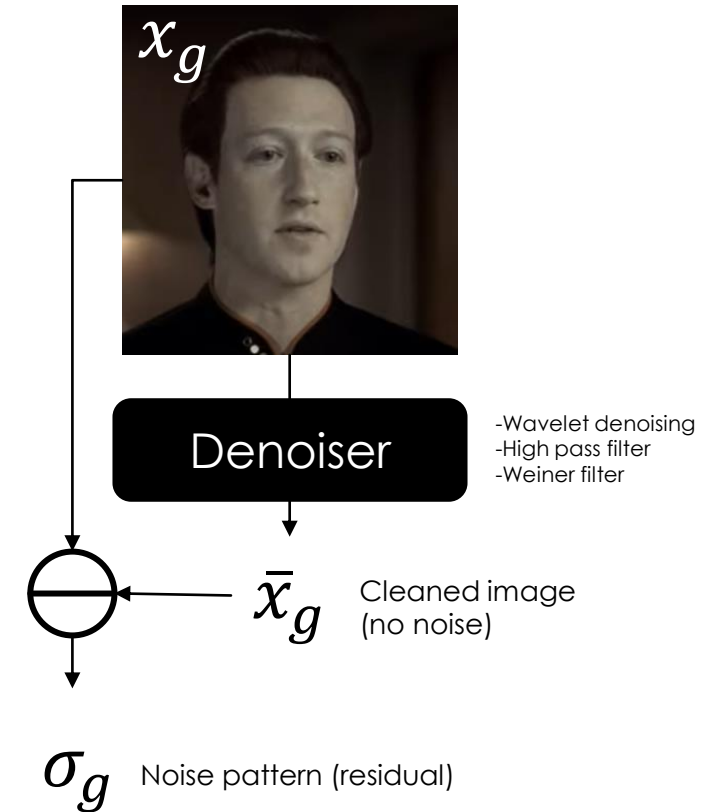
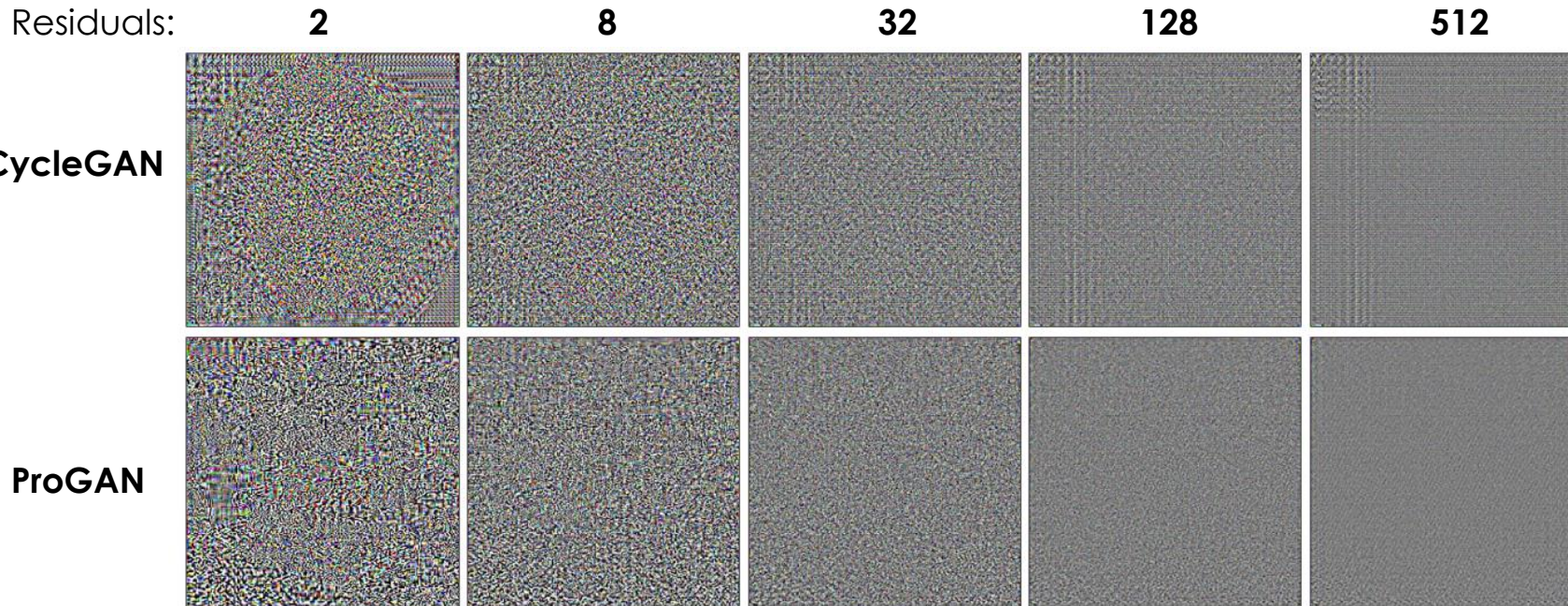
Ciftci U, et al. FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals. 2020

# Deepfake Defences

## 2. Directed Approaches

### 2.3 Spatial – Forensics

GANs have Fingerprints!



# Deepfake Defences

## Prevention

1. Data Provenance



3. Cyber Security



2. Counter Attacks



4. Awareness





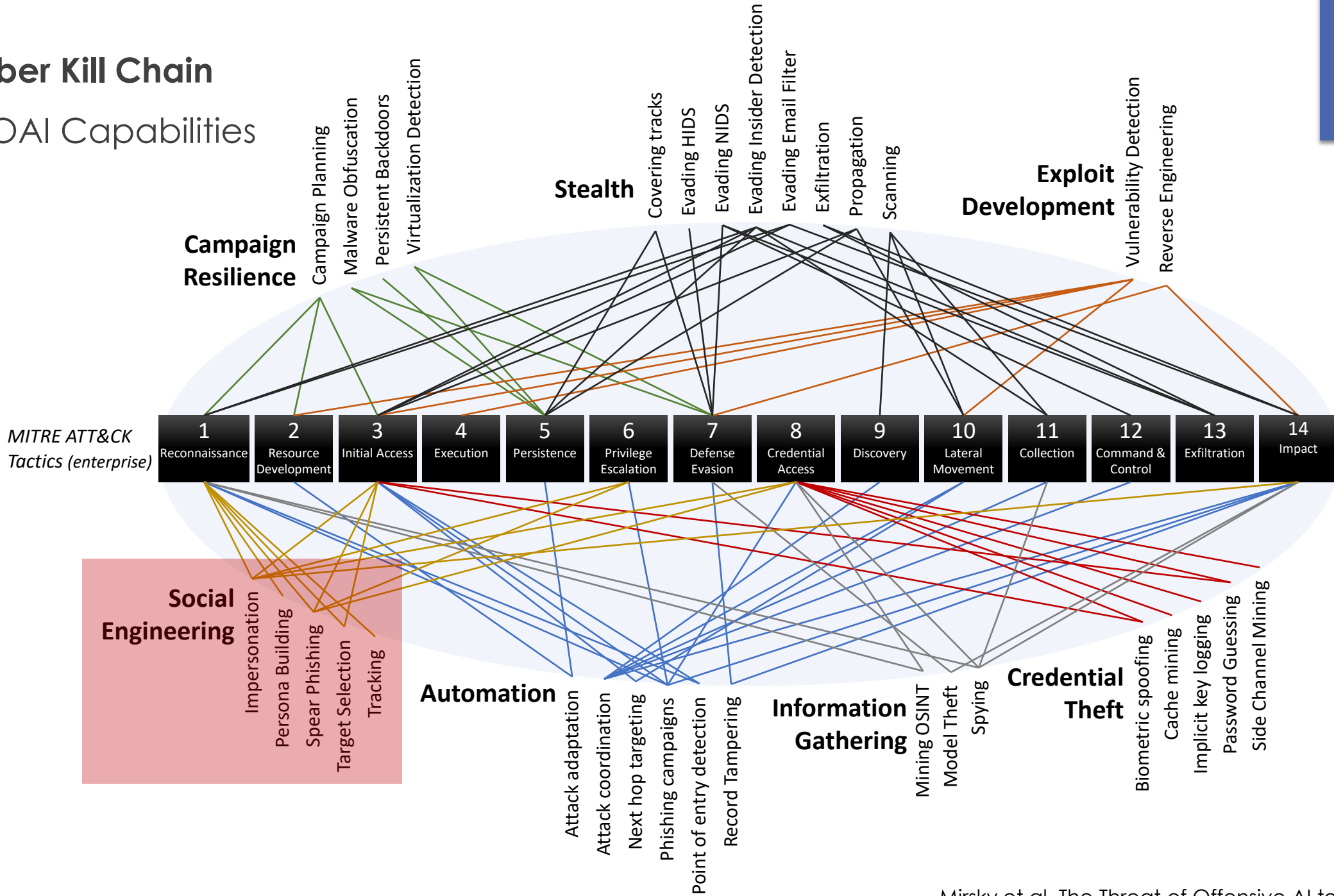
# The Threat Horizon





# Cyber Kill Chain

## 32 OAI Capabilities

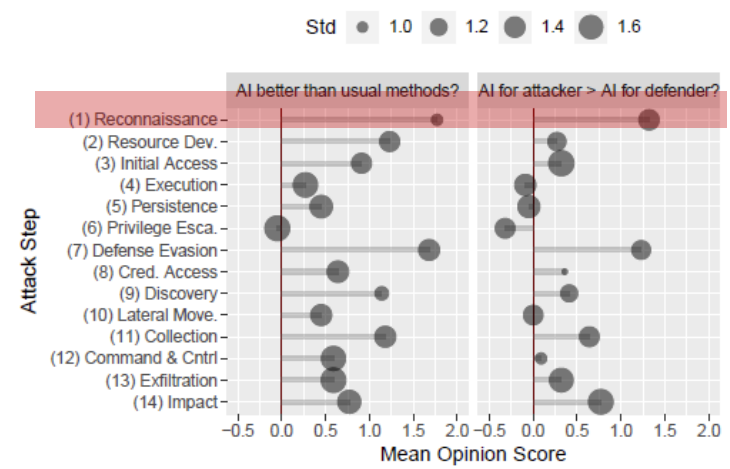
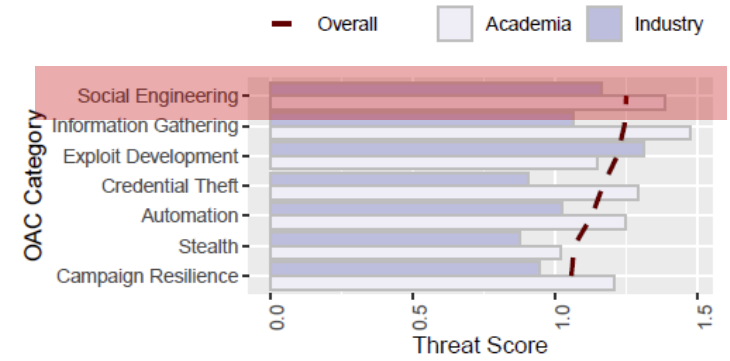
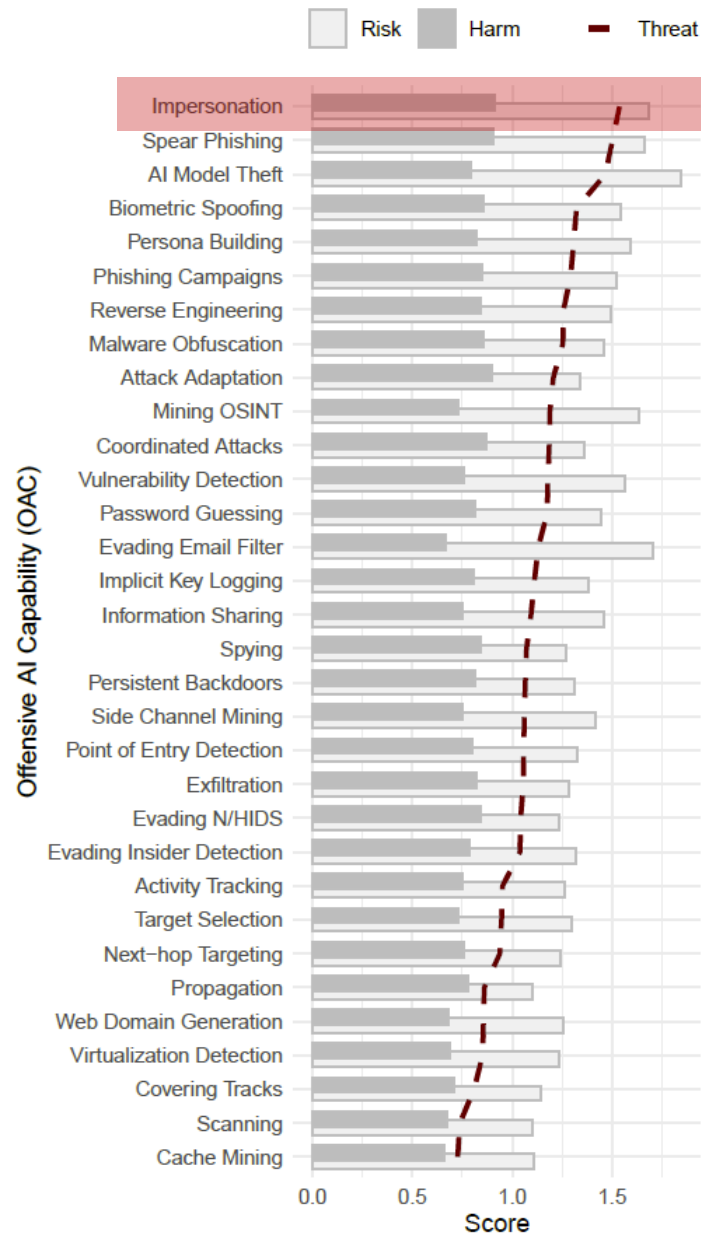


# Threat Horizon

## Survey Results

35 Industry+Academia

Top: SE – easy to achieve, most harm, hard to defeat, ...



# Threat Horizon

What are they so afraid of?

**Real Time Deepfakes**



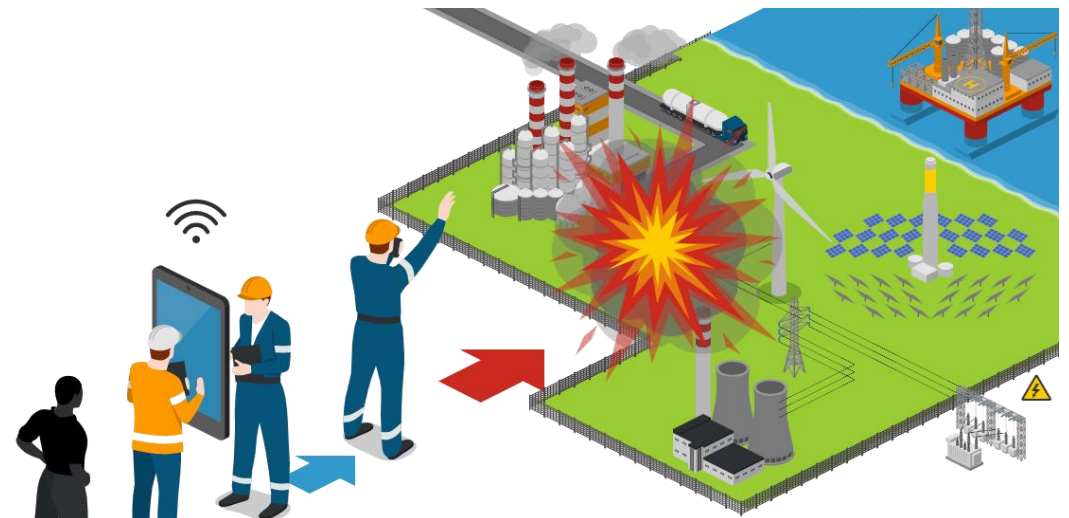
# Threat Horizon

## The Threat To Companies

Bypasses all defences...



## The Threat To Critical Infrastructure



Get Information  
Change setting  
Turn on/off system  
...

# Threat Horizon

## The Threat To Individuals

### Example RT-DF attack:

1. Find connection on facebook between A and B
2. Call A and clone voice from 3 second chat
3. Call B with A's voice, read script

Automate the campaign...  
(could replace email phishing)



“Mom, I need help!”



# Threat Horizon

## Real Time Deepfakes

### Re-enactment



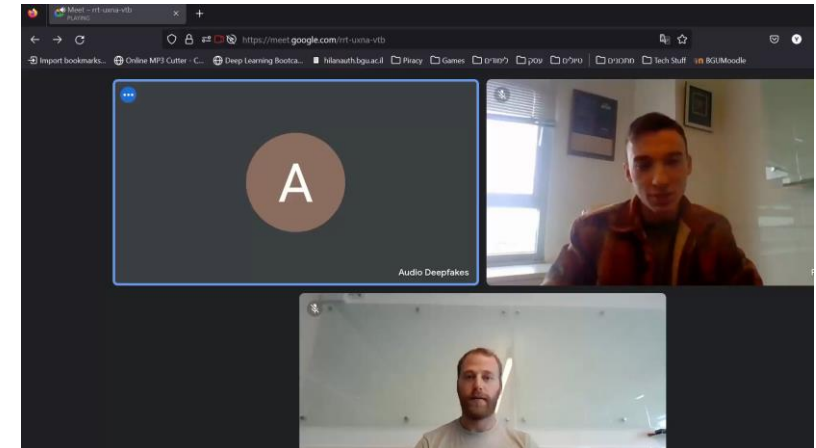
Avatartify - First order motion model for image animation. NeurlPs

### Replacement



Metaphysics AGT 2022

### Voice Cloning



StarGAN-VC – our online version

Re-enactment of a single photo!

# What Can We Do?

- ▶ Existing solutions look for artifacts...
- ▶ But the quality of DFs will only get better!

**Answer 1:** Better Data provenance

**Answer 2:** Awareness

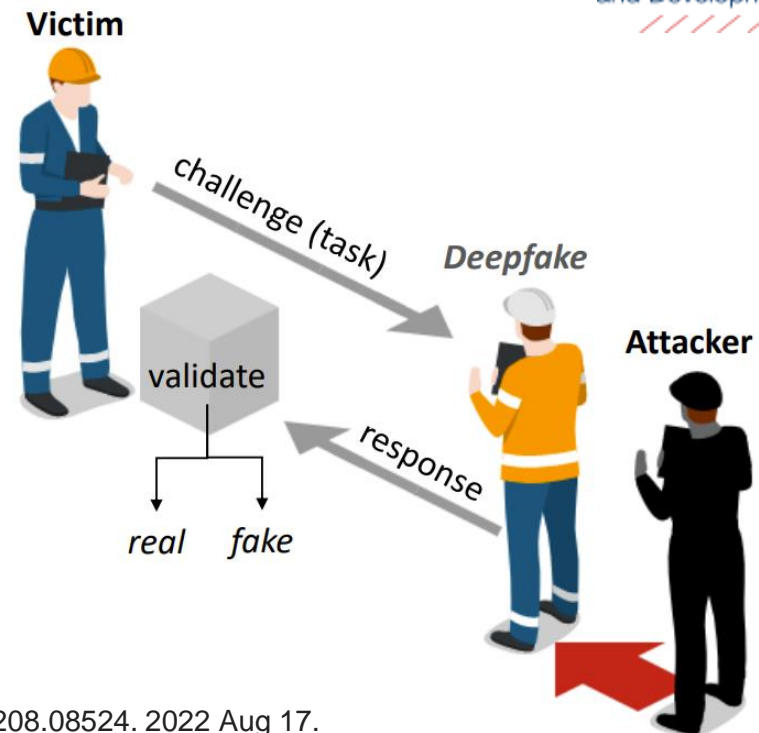
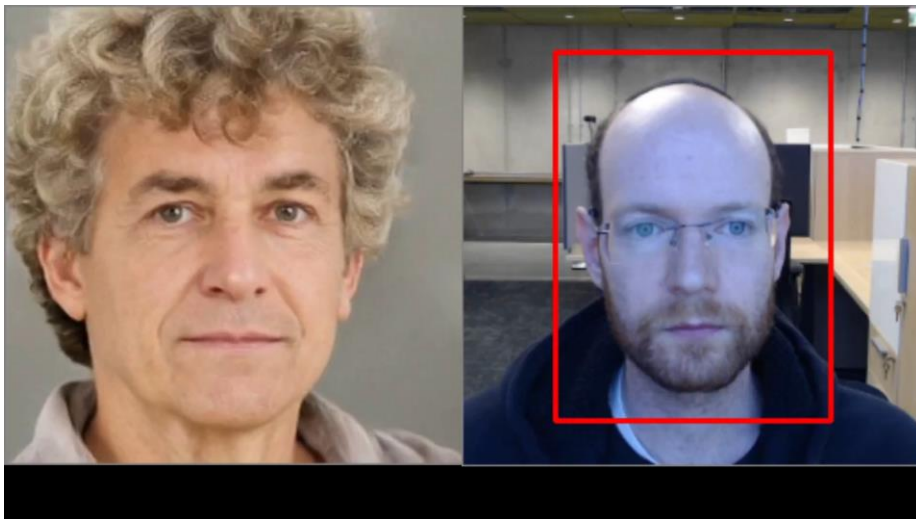
**Answer 3:** Counter Attacks

# DF-CAPTHCA

## A Turing Test on content generation



DFs breakdown when pushed beyond their design/capabilities



# DF-CAPTCHA

Looking for a Commercialization Partner!

## Potential Products:

- ▶ Virtual Meeting Authentication
  - ▶ **Service/feature:** validate participants in waiting room
- ▶ Smartphone Call Screening
  - ▶ **App:** automatically validate incoming calls
- ▶ Policy Enforcement
  - ▶ **App:** forward calls that seem suspicious



# Thank you!

Dr. Yisroel Mirsky

*Head of the Offensive AI Research Lab*

yisroel@bgu.ac.il

<https://offensive-ai-lab.github.io/>

