



# SEI-CMU Deepfakes Work

Catherine Bernaciak

Dominic Ross

Matthew Walsh

Software Engineering Institute  
Carnegie Mellon University  
Pittsburgh, PA 15213

Copyright 2023 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](mailto:permission@sei.cmu.edu).

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

DM23-0381

# Overview

- Introduction
- What is a deepfake?
- How are deepfakes created?
- How are deepfakes detected?
- Outlook

# Machine learning methods allow creation of images that are entertaining, funny



## This Person Does Not Exist

The site that started it all, with the name that says it all. Created using a style-based generative adversarial network (StyleGAN), this website had the tech community buzzing with excitement and intrigue and inspired many more sites.



## This Foot Does Not Exist

Note that this is an SMS chatbot. You can text it to get pictures of feet. The pictures are animated. The feet are nonexistent. Why would you want to do this? Who knows.

Created by MSCHF.



## This Cat Does Not Exist

These purr-fect GAN-made cats will freshen your feeline-gs and make you wish you could reach through your screen and cuddle them. Once in a while the cats have visual deformities due to imperfections in the model – beware, they can cause nightmares.



## This Artwork Does Not Exist

Be inspired by minimalism, realism, post-modernism, pre-modernism, modernism, and ancientism (not actually a thing). No matter your art preferences, you can find it here with enough refreshing.

Created by Michael Friesen.



## This Rental Does Not Exist

Why bother trying to look for the perfect home when you can create one instead? Just find a listing you like, buy some land, build it, and then enjoy the rest of your life.



## This Chemical Does Not Exist

Who said drug discovery is hard? Just refresh until you find the right chemical. In all seriousness, the fact that this renders a 3D model with the correct bond pairings is impressive.

<https://thisxdoesnotexist.com>

## ...and a little scary

# Generation of artificial faces is one area that has gained much attention

Picture A



Picture B



Can you spot the fake?

"Well, as with anything, if you want to believe you can find reasons to."

Google

Find image source

Visual matches

CONAN

youtube.com

When Tom Hanks Read For "Splash" | CONAN o...

Not everything can be a reverse image search...

# There are many reasons for artificially manipulating faces

- Accessibility
- Education
- Art and demonstration
- Satire/Humor
- Editing and special effects

- Identity theft
- Fraud
- Sale of non-consensual content
- Personal attack
- Misinformation, disinformation, and mal-information\*



Some are problematic

# The threats posed by deepfakes are growing

- Open source software like Faceswap and DeepFaceLab are publicly available
- Modest hardware requirements
- No code / low code options to generate deepfakes exist
- New methods reduce need for post-processing

The screenshot shows a GitHub search results page for the query 'deepfakes/faceswap'. The results are as follows:

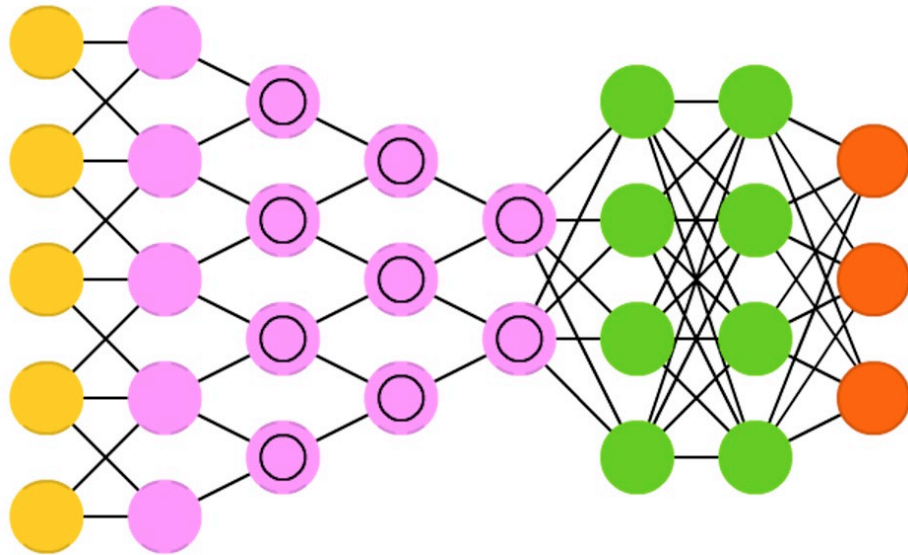
- deepfakes/faceswap** (Sponsor): Deepfakes Software For All. 43.2k stars, Python, GPL-3.0 license, updated 8 hours ago. Tags: machine-learning, deep-learning, deep-neural-networks, faceswap, neural-networks, face-swap, deeplearning, neural-nets, deepface, deepfakes, fakeapp, deep-face-swap, openfaceswap, myfakeapp.
- iperov/DeepFaceLab**: DeepFaceLab is the leading software for creating deepfakes. 36.7k stars, Python, GPL-3.0 license, updated 3 weeks ago. Tags: machine-learning, deep-neural-networks, deep-learning, faceswap, neural-networks, face-swap, deeplearning, arxiv, neural-nets, deepface, deepfakes, fakeapp, deep-face-swap, deepfacelab, creating-deepfakes.
- yuanxiaosc/DeepNude-an-Image-to-Image-technology** (Sponsor): DeepNude's algorithm and general image generation theory and practice research, including pix2pix, CycleGAN, UGATIT, ... 4.3k stars, Python, updated on Oct 2, 2021. Tags: dcgan, vae, image-generation, pix2pix, image-to-image, nerual-style, cycle-gan, deepface, deepfakes, tensorflow2, style-gan, deepnude, zao, sin-gan.
- joshua-wu/deepfakes\_faceswap**: from deekfakes' faceswap: <https://www.reddit.com/user/deepfakes/> 3k stars, Python, updated on Feb 2, 2018.
- sensity-ai/dot**: The Deepfake Offensive Toolkit. 2.9k stars, Python, BSD-3-Clause license, updated 3 weeks ago.

It is becoming harder to detect deepfakes than to create them

# Overview

- Introduction
- **What is a deepfake?**
- How are deepfakes created?
- How are deepfakes detected?
- Outlook

# What is a deepfake?



A deep neural network (DNN) is a neural network with many layers used to convert inputs to targets

A deepfake is a media file of a human subject that has been manipulated using a DNN

**Deepfakes are more complex than other digital forgeries**

# Evolution of deepfakes

## General Developments

Generative Adversarial Neural Nets<sup>4</sup>

Convolutional Neural Nets  
for edge detection<sup>1</sup>

Nvidia CUDA<sup>2</sup>

Imagenet<sup>3</sup>

Neural Radiance Fields<sup>5</sup>

1995

2000

2005

2010

2015

2020

Facial animation systems used for  
video re-writes<sup>6</sup>

DeepFace<sup>7</sup>

DeepId<sup>8</sup>

CycleGAN<sup>11</sup>

Metahuman/DF<sup>13</sup>

DeepFaceLab<sup>12</sup>

Faceswap<sup>10</sup>

pix2pix<sup>9</sup>

## Face-specific Applications

## Deepfakes reddit forum

# Summary of architectures

Architecture	Strengths	Weaknesses
Autoencoder	<ul style="list-style-type: none"><li>• High quality 2D outputs</li><li>• Target individuals</li><li>• Mature toolboxes exist</li></ul>	<ul style="list-style-type: none"><li>• Time to train</li><li>• Require target footage</li><li>• Post-production needs</li></ul>
GAN	<ul style="list-style-type: none"><li>• Highest quality 2D outputs</li><li>• Synthesize whole faces</li><li>• Mature toolboxes exist</li></ul>	<ul style="list-style-type: none"><li>• Produce distinct artifacts</li><li>• Less suitable for targeting specific individuals</li><li>• Less suitable for generating videos</li></ul>
NeRF	<ul style="list-style-type: none"><li>• High quality 3D outputs</li><li>• Requires less data</li></ul>	<ul style="list-style-type: none"><li>• Less advanced research</li></ul>
DF Live	<ul style="list-style-type: none"><li>• High quality 3D outputs</li><li>• Requires less data</li><li>• Real-time fakes</li></ul>	<ul style="list-style-type: none"><li>• Less advanced research</li></ul>

# Overview

- Introduction
- What is a deepfake?
- **How are deepfakes created?**
- How are deepfakes detected?
- Outlook

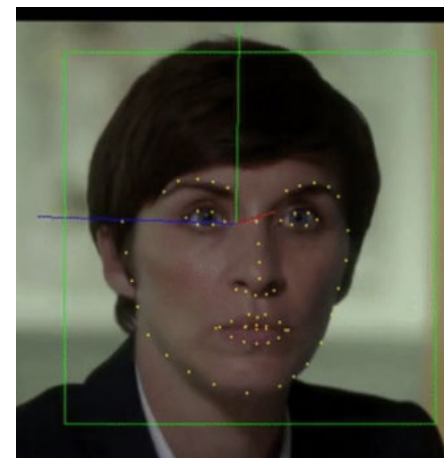
# How are deepfakes created with 5 steps?

## 1. Gather source & destination video (CPU)

- High-quality (4K), voluminous (> 10 minutes) source footage and destination footage of subject with similar appearance

## 2. Extraction (CPU/GPU))

- Faces isolated from each frame using DNN & ML based facial recognition models (S3FD)
  - a. faces detected in frame
  - b. faces aligned, facial landmarks identified
  - c. mask of face segmented from frame



# How are deepfakes created with 5 steps?

## 3. Training (GPU)

- DFL and FS mostly use autoencoder-decoder networks with CNN layers
- FS and DFL ~  $O(10)$  models each, some very similar

## 4. Conversion (CPU/GPU)

## 5. Post-processing (CPU)

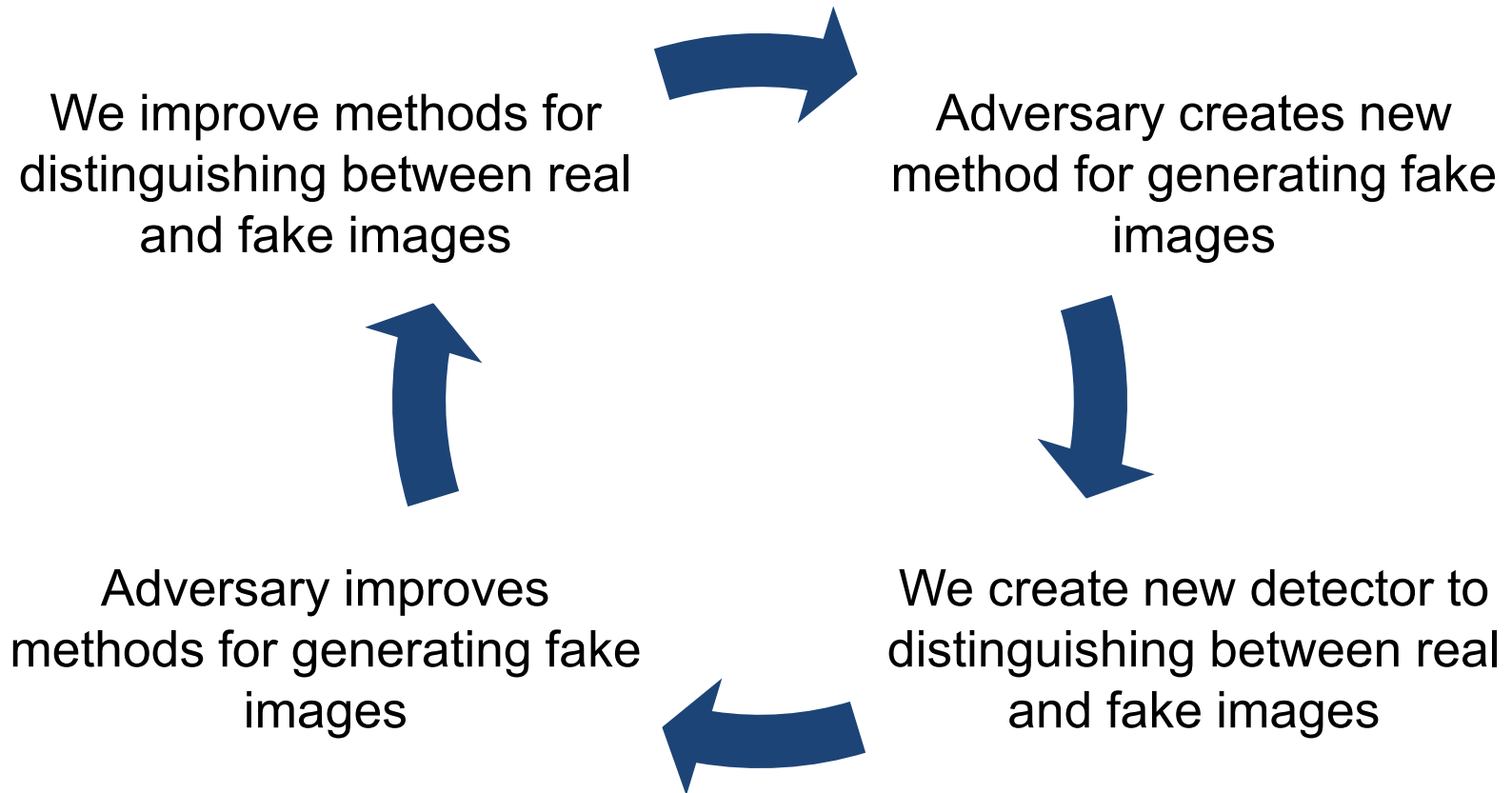
# Overview

- Introduction
- What is a deepfake?
- How are deepfakes created?
- **How are deepfakes detected?**
- Outlook

# Deepfake Detection

- Machine-augmented detection is needed to deal with increased volume and complexity of deepfakes
- Deepfake detectors *discriminate* between real and *fake* images
- To develop models that automatically discriminate:
  - **Humans** may feed model features associated with differences
  - **Computers** may 'learn' useful features on their own
  - The two approaches may be combined

# Generation and Detection Loop



**This is a costly game of inches**

# Human Determined Features

What *describable* features make deepfakes different?

- ‘Obvious’ errors (e.g. two heads in GANs)
- Facial boundaries blurring into background
- Asymmetries (e.g. earrings not matching)
- Inconsistent light sources
- Odd color frequencies
- Irregular ‘heartbeats’
- Discontinuity between frames
- Lack of variation

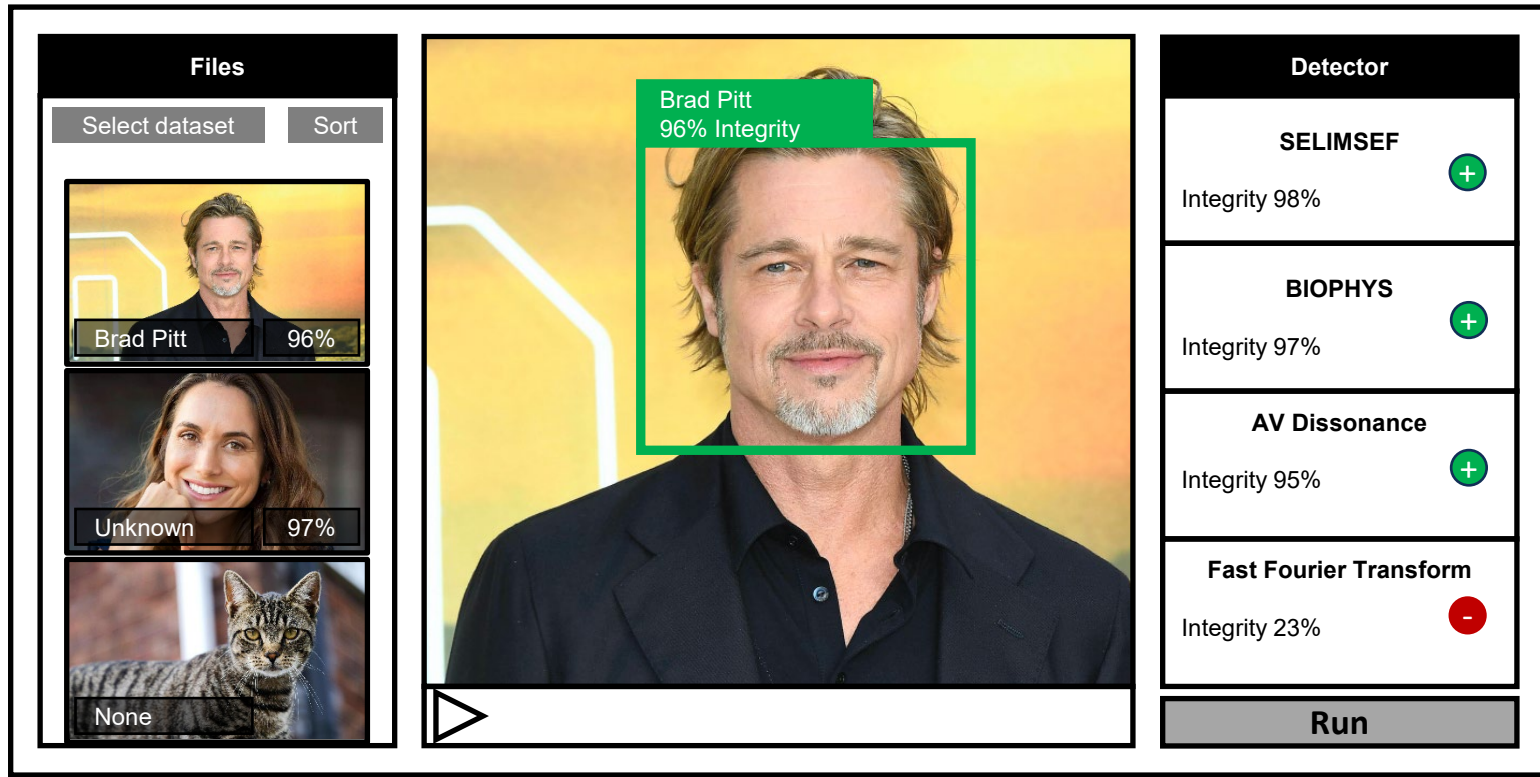
# SEI Deepfake Detection Work

SEI is building a deepfake detection tool that combines different detection modes to give an ensemble assessment of authenticity:

Information source	Derived features	Detection Mode
Pixels	Spatial features	Distinct spatial features
Eyes	Eyeblink measurement	Reduced blink rate
Facial features	Emotion classification	More neutral emotions
Color changes	Heart rate	Greater temporal variability
Frequency domain	Spectrogram	Less power at high frequencies
Digital noise	Pixel-response non-uniformity	Inconsistent PRNU
Audio	Audio-visual dissonance	Greater dissonance

# SEI Deepfake Detection Tool

- Combining different detection modes into one framework
- Framework is extensible – different detection methods can be added
- The tool is usable by an analyst, does not require algorithm expertise
- GUI interface similar to Medifor



# Final Thoughts

- It is difficult, but not impossible to make high-quality deepfakes with enough time, effort, skill, and resources
- Given the potential scale and complexity of deepfakes, automatic detection methods are needed
- It is difficult, but not impossible to detect deepfakes automatically
- However, detection methods must continue to evolve to maintain upper-hand