SEI-CMU Deepfakes Work

Catherine Bernaciak

Dominic Ross

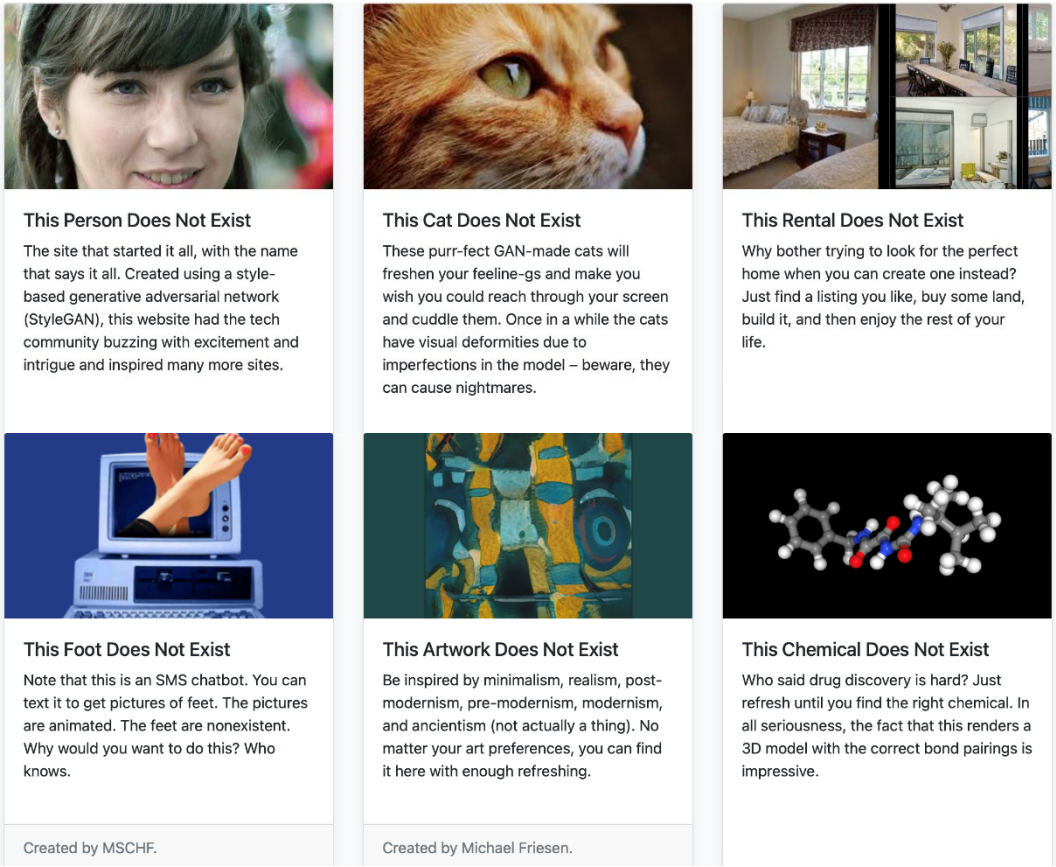
Matthew Walsh

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

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



https://thisxdoesnotexist.com

…and a little scary

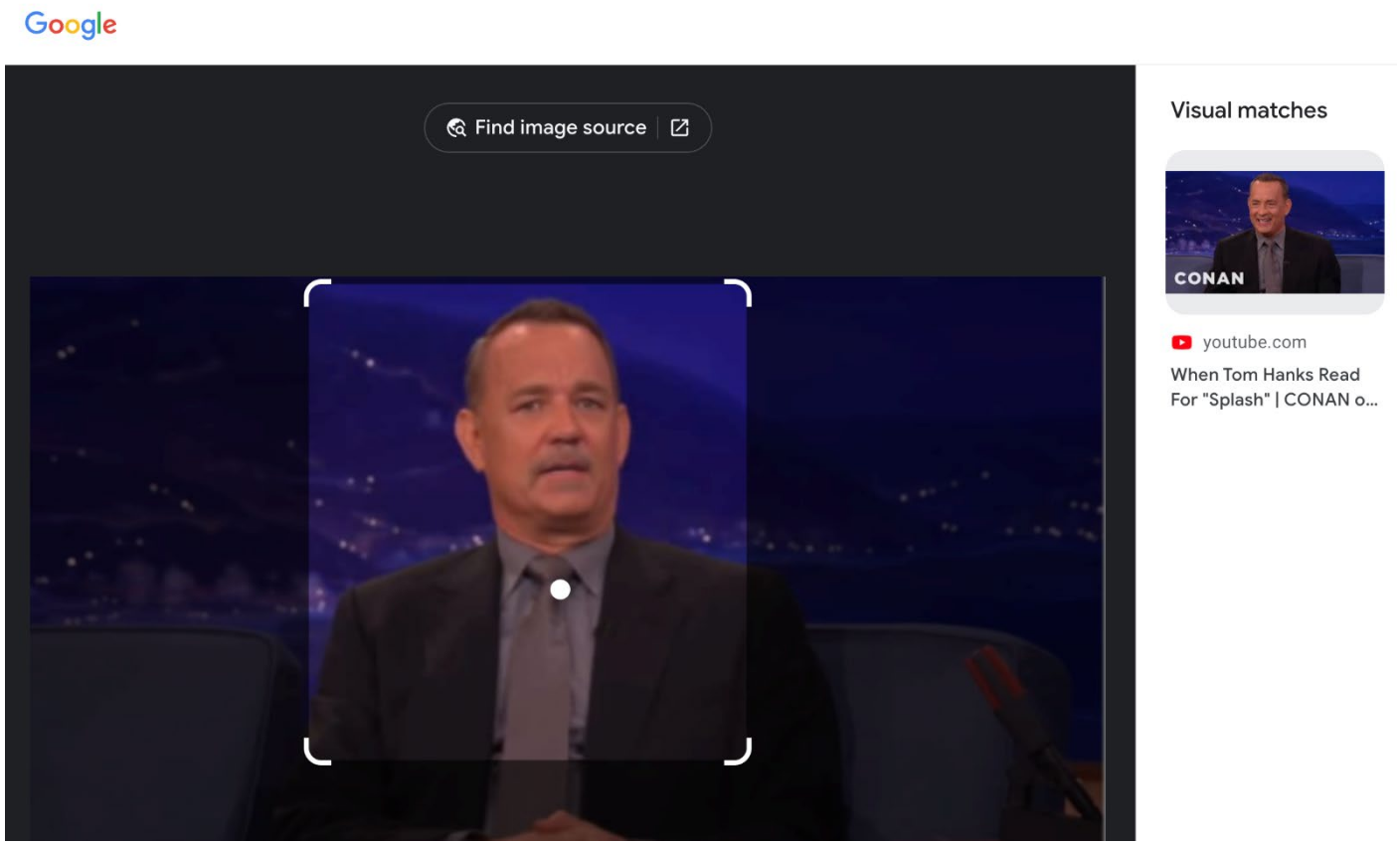
Generation of artificial faces is one area that has gained much attention

Picture A

Picture B

Can you spot the fake?

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



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*

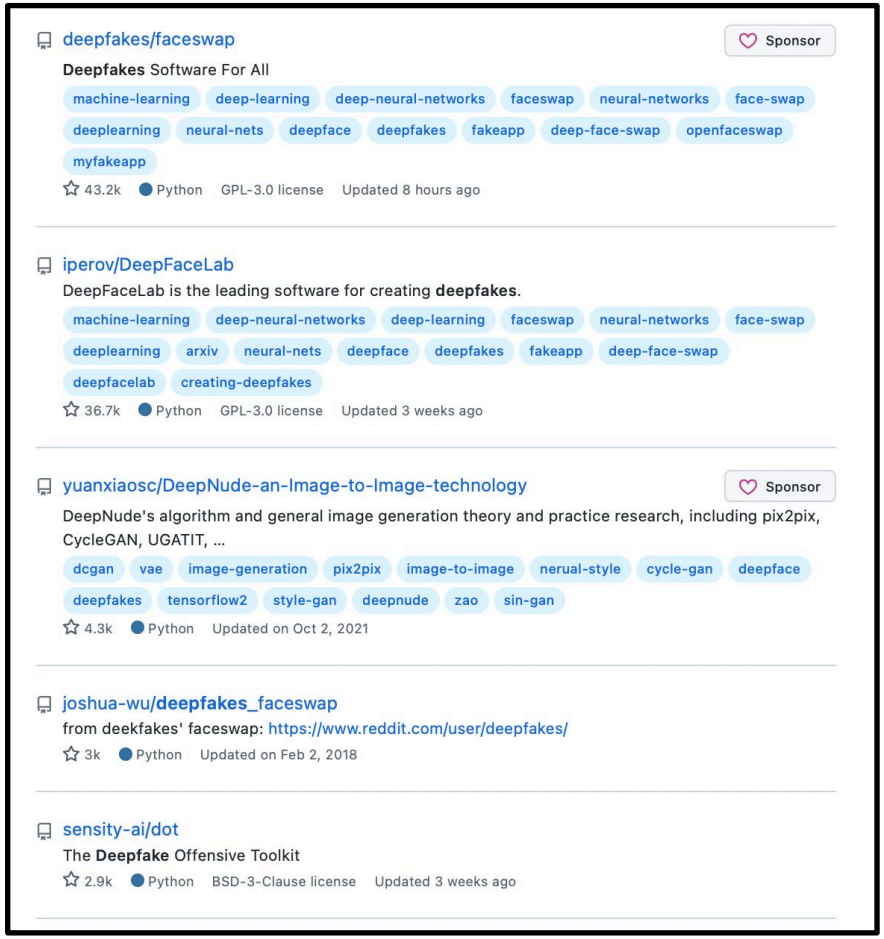
No **Yes**

Intent to mislead and/or cause harm?

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

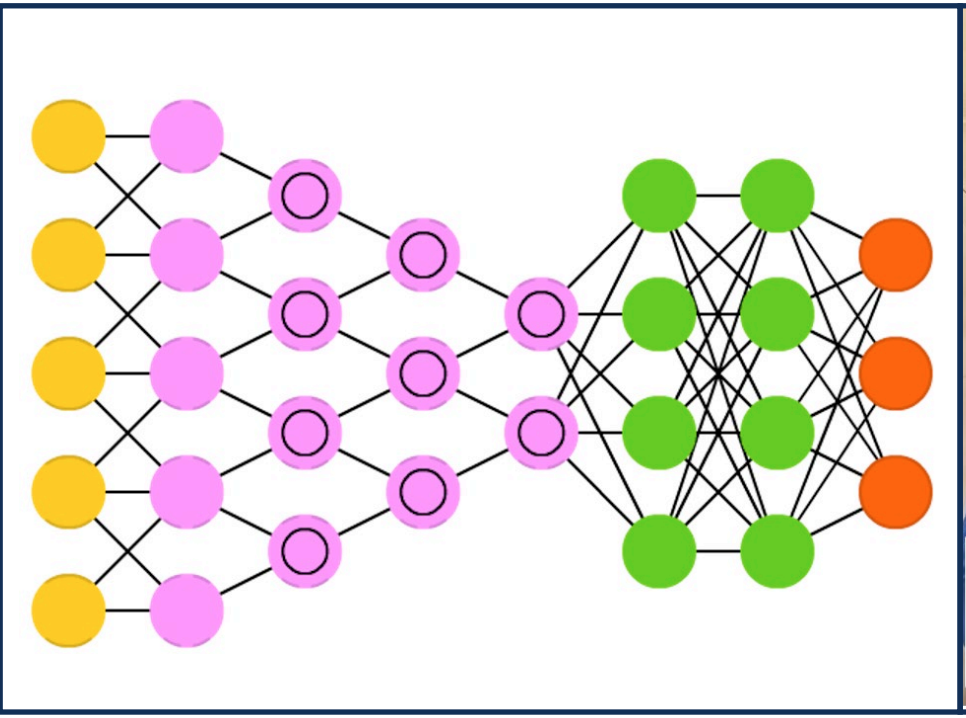


It is becoming harder to detect deepfakes than to create them

Overview

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

What is a deepfake?

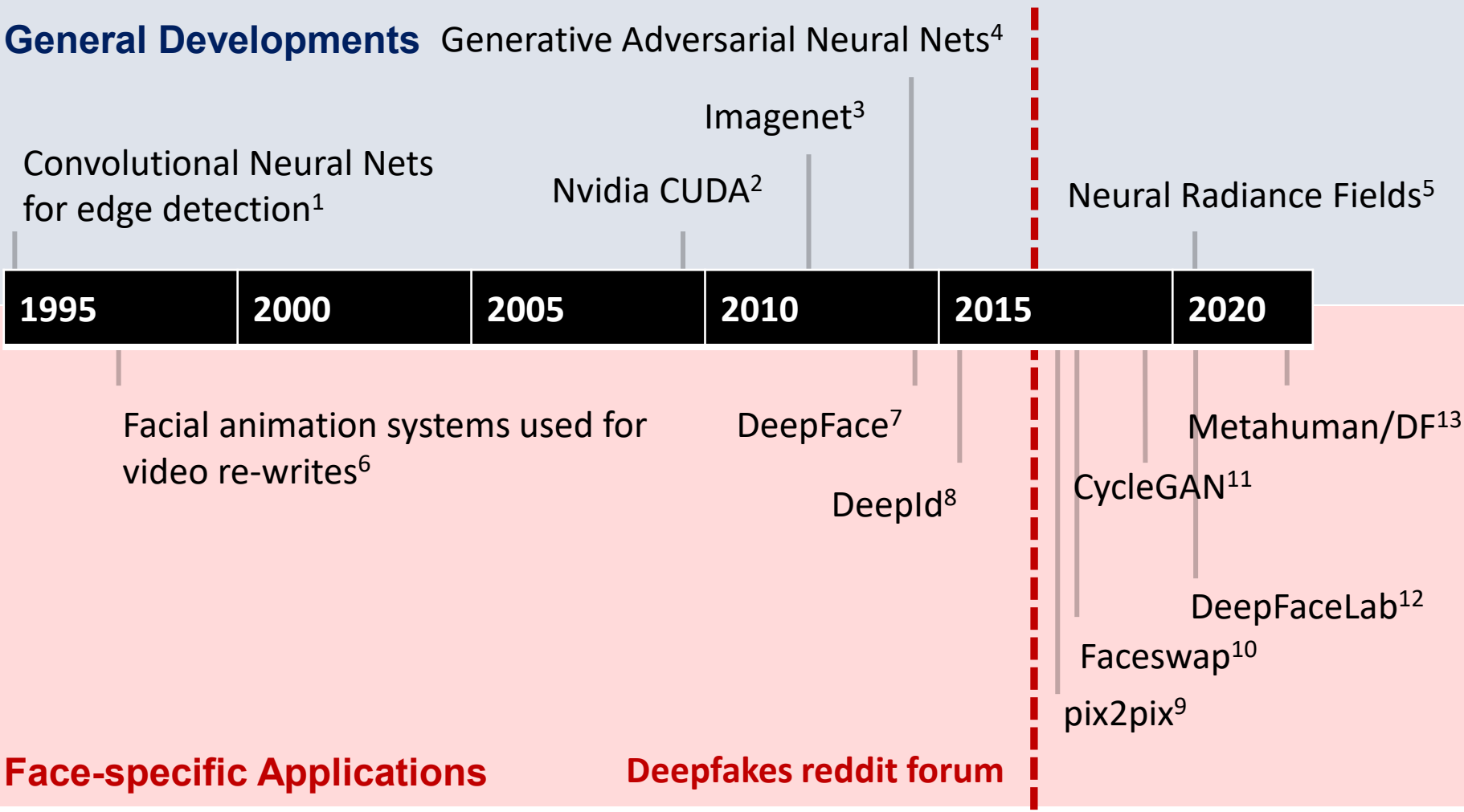


A deep neural network (DNN) is a neural network with many layers used to convert inputs to targets

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



Summary of architectures

Architecture	Strengths	Weaknesses
Autoencoder	• High quality 2D outputs • Target individuals • Mature toolboxes exist	• Time to train • Require target footage • Post-production needs
GAN	• Highest quality 2D outputs • Synthesize whole faces • Mature toolboxes exist	• Produce distinct artifacts • Less suitable for targeting specific individuals • Less suitable for generating videos
NeRF	• High quality 3D outputs • Requires less data	• Less advanced research
DF Live	• High quality 3D outputs • Requires less data • Real-time fakes	• Less advanced research

Overview

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

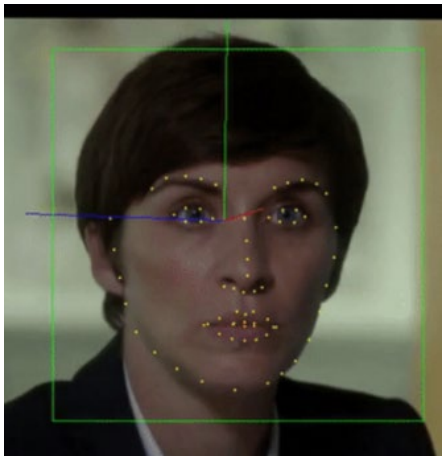
How are deepfakes created with 5 steps?

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)

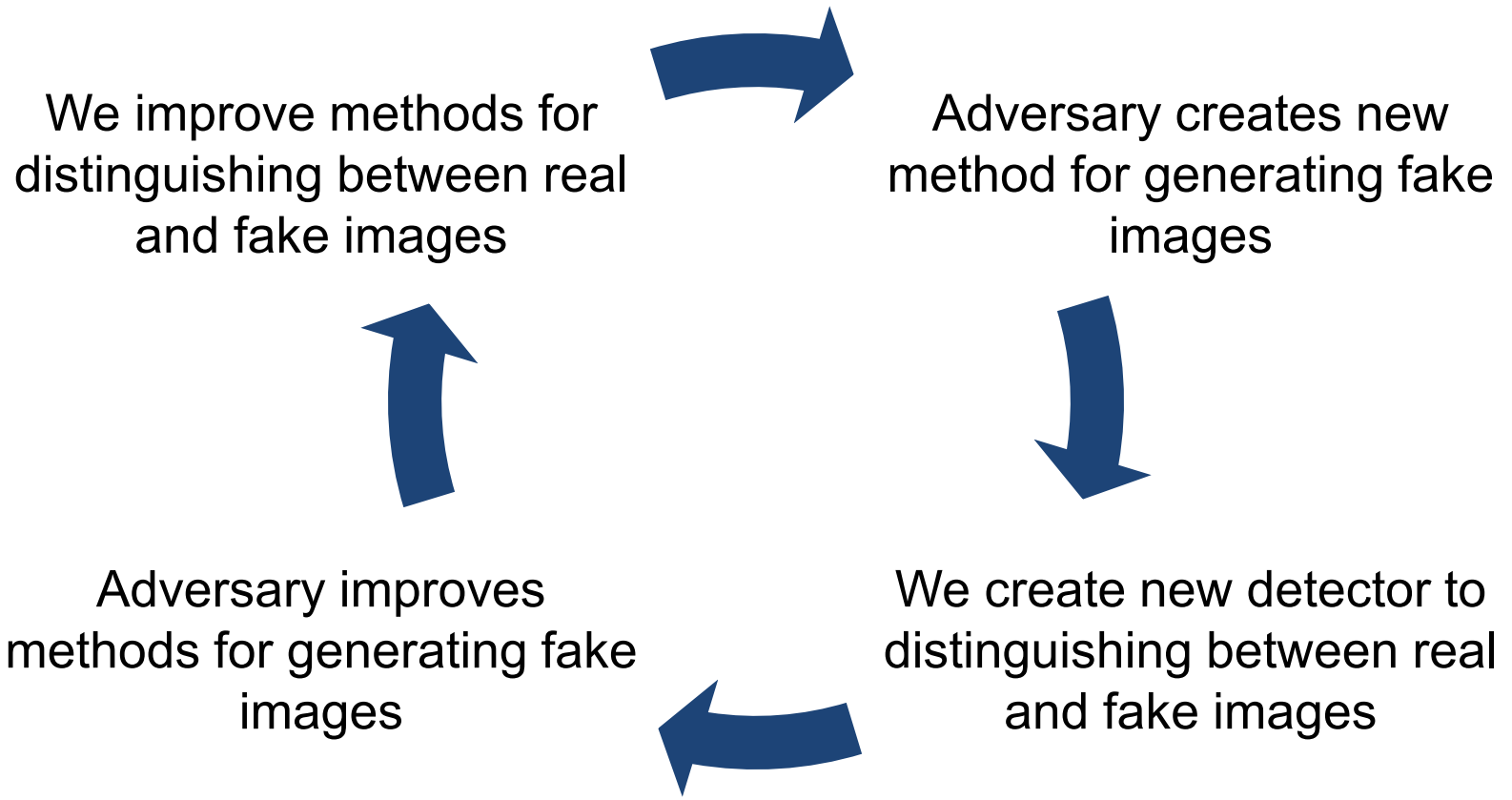
Overview

- 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

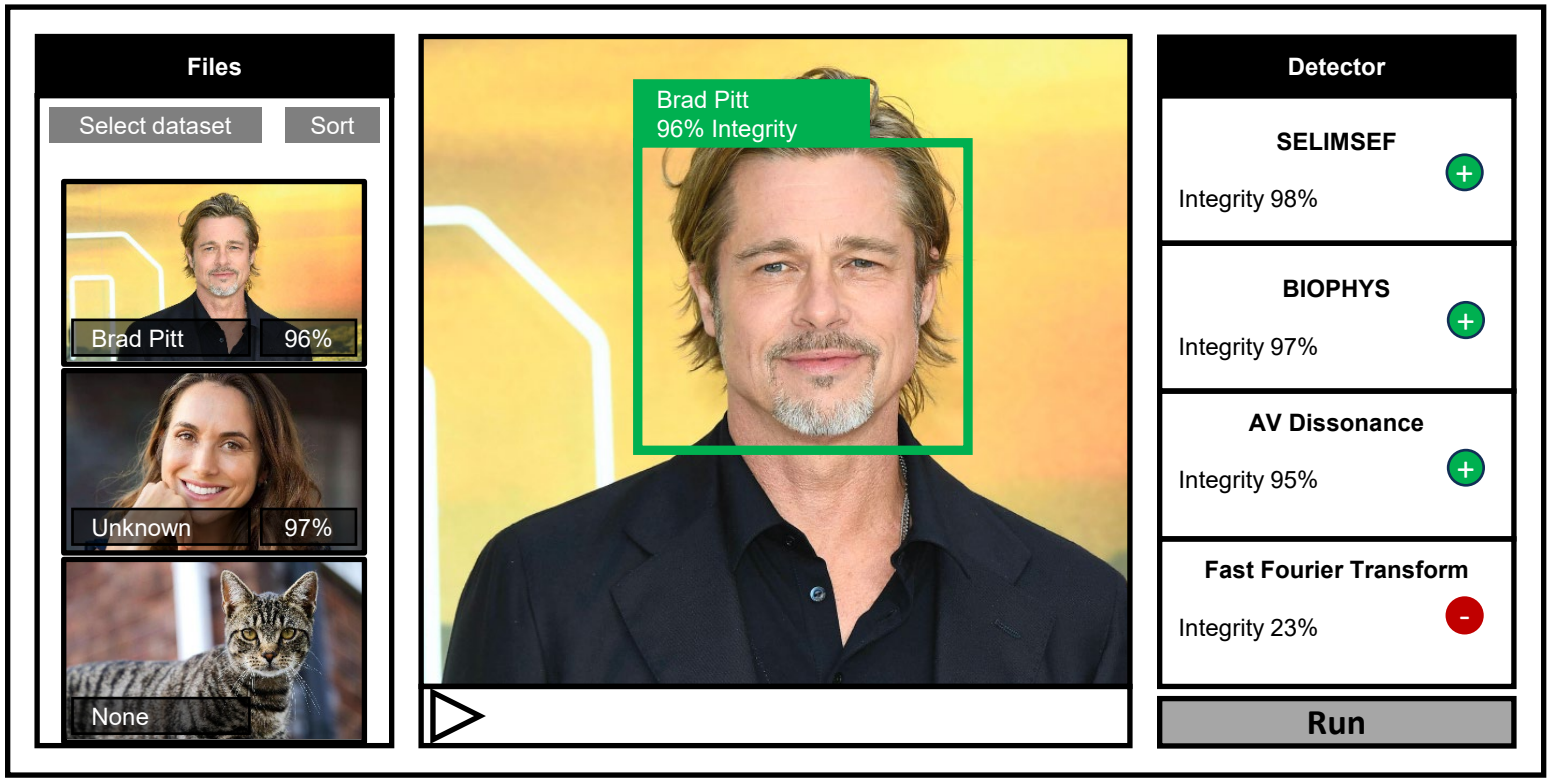
SEI Deepfake Detection Work

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