OWASP™

DEEPFAKES: A GROWING CYBERSECURITY CONCERN

What Are Deepfakes?

A combination of "deep learning" and "fake", deepfakes are hyper-realistic videos digitally manipulated to depict people saying and doing things that never actually happened.

Deepfakes rely on neural networks that analyze large sets of data samples to learn to mimic a person's facial expressions, mannerisms, voice, and inflections.

The process involves feeding footage of two people into a deep learning algorithm to train it to swap faces. In other words, deepfakes use facial mapping technology and AI that swaps the face of a person on a video into the face of another person.

What Are Deepfakes?

Deepfakes are difficult to detect, as they use real footage, can have authentic-sounding audio, and are optimized to spread on social media quickly. Thus, many viewers assume that the video they are looking at is genuine.

Deepfakes target social media platforms, where conspiracies, rumors, and misinformation spread easily, as users tend to go with the crowd.

At the same time, an ongoing ‘infopocalypse’ pushes people to think they cannot trust any information unless it comes from their social networks, including family members, close friends or relatives, and supports the opinions they already hold.

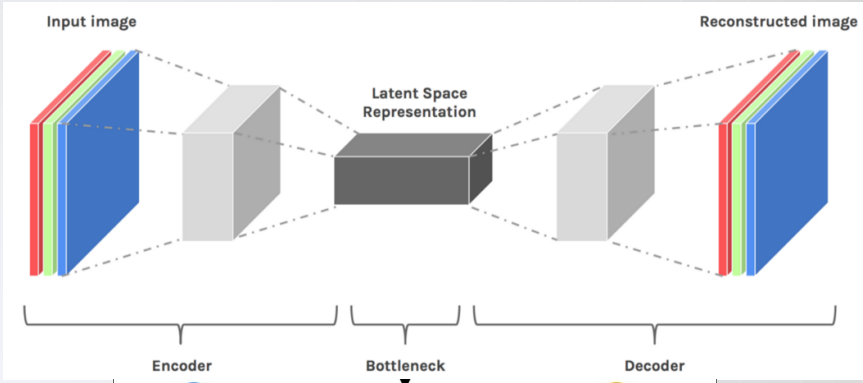
In fact, many people are open to anything that confirms their existing views even if they suspect it may be fake.

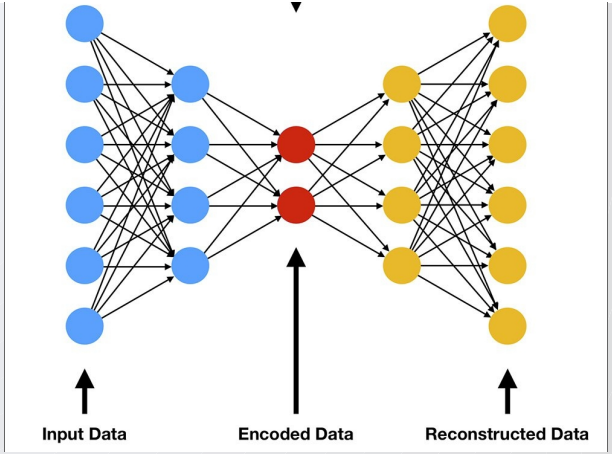
How Do Deepfakes Work?

An Autoencoder is a special type of neural network that is trained to copy its input to its output.

For example, given an image of a handwritten digit, an autoencoder first encodes the image into a lower dimensional latent representation, then decodes the latent representation back to an image.

An autoencoder learns to compress the data while minimizing the reconstruction error.





How Do Deepfakes Work?

The result is that the two autoencoders have a shared encoder that can "read" either a Mark Zuckerberg face or a Mr. Data face.

The goal is for the encoder to use the same representation for things like head angle or eyebrow position whether it's given a photo of Mark Zuckerberg or a photo of Mr. Data.

And that, in turn, means that once you've compressed a face with the encoder, you can expand it using either decoder.

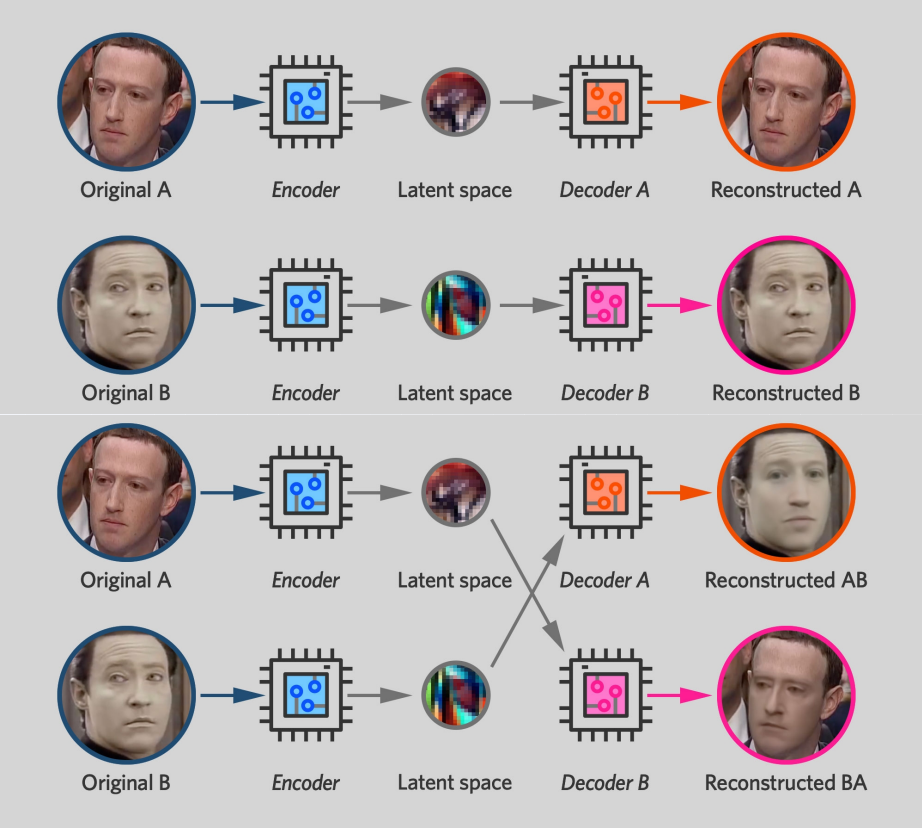
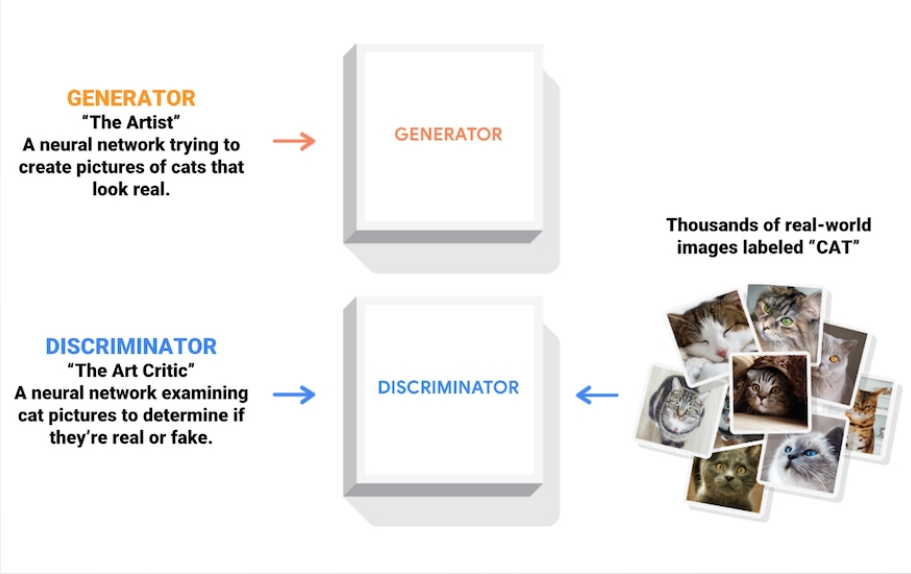


Photo Credit: Aurich Lawson - https://arstechnica.com/science/2019/12/how-i-created-a-deepfake-of-mark-zuckerberg-and-star-treks-data/

How Do Deepfakes Work?

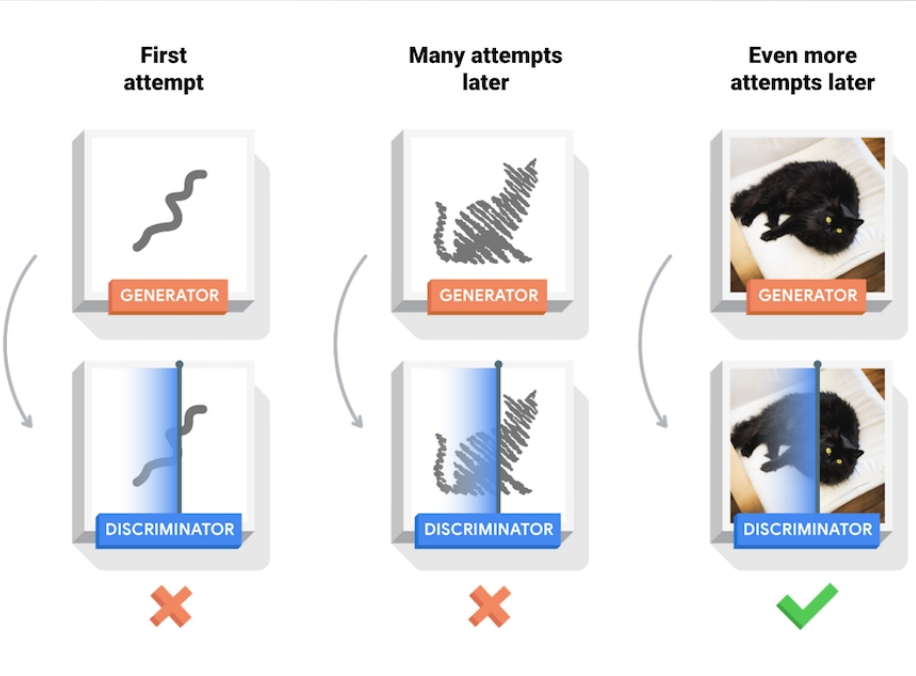
Generative Adversarial Networks (GANs) are one of the most interesting ideas in computer science today. Two models are trained simultaneously by an adversarial process. A generator ("the artist") learns to create images that look real, while a discriminator ("the art critic") learns to tell real images apart from fakes.



How Do Deepfakes Work?

During training, the generator progressively becomes better at creating images that look real, while the discriminator becomes better at telling them apart.

The process reaches equilibrium when the discriminator can no longer distinguish real images from fakes.



Deepfake Examples

+ Tech Timeline

DEEPFAKE TIMELINE

2016 NOV

Highlights:

- Adobe #VoCo is an audio manipulator that allows you to change words in a voiceover simply by typing new words. Presented live during the Adobe MAX 2016

https://www.youtube.com/watch?v=I3l4XLZ59iw

2018 FEB

Platforms Banning Deepfakes:

- Several websites, including Discord, Gfycat, and Twitter, ban deepfakes with varying degrees of success.

2018 APR

Obama Deepfake Video:

- Deepfake video of former US president Barack Obama raises mainstream awareness.

https://www.youtube.com/watch?v=cQ54GDm1eL0

2018 MAY

US Senator Voices His Concern About Deepfakes:

- US Senator Marco Rubio voices his concerns aboutdeepfakes at the Senate Intelligence Committee nomination hearing

“#VoCo. Adobe’s Audio Manipulator” Demo

https://www.youtube.com/watch?v=I3l4XLZ59iw

Obama Deepfake Video

https://www.youtube.com/watch?v=cQ54GDm1eL0

DEEPFAKE TIMELINE

2019 APR

Highlights:

David Beckham Deepfake Video:

- The recent global campaign showing Malaria survivors speaking through David Beckham to help raise awareness around the Malaria Must Die initiative spooked a lot of people

2019 JUN

House holds hearing on "deepfakes" and artificial intelligence amid national security concerns:

- The House Intelligence Committee heard from experts on the threats that so-called "deep fake" videos and other types of artificial intelligence-generated synthetic data pose to the U.S. election system and national security at large.

2019 JUL

FaceApp:

- AI photo editor FaceApp goes viral after adding AI-based age filter.

2019 AUG

DARPA Is Taking On the Deepfake Problem:

- The Defense Department is looking to build tools that can quickly detect deepfakes and other manipulated media amid the growing threat of “large-scale, automated disinformation attacks.”

David Beckham Deepfake Video

https://www.youtube.com/watch?v=QiiSAvKJIHo

DEEPFAKE TIMELINE

Highlights:

2019 SEP

CEO Deepfake Scam:

- Criminals used artificial intelligence-based software to impersonate a chief executive's voice and demand a fraudulent transfer of €220,000 ($243,000) in what cybercrime experts described as an unusual case of artificial intelligence being used in hacking.

2019 OCT

California and Texas ban political deepfake videos:

- California and Texas have passed a law meant to prevent altered "deepfake" videos from influencing elections in a plan that has raised free speech concerns..

2020 DEC

Deepfake Queen: 2020 Alternative Christmas Message:

- An alternative Christmas message for a very alternative year.

https://www.youtube.com/watch?v=IvY-Abd2FfM

2021 JUL

DeepFaceLive:

- Real-time face swap for PC streaming or video calls.

https://github.com/iperov/DeepFaceLive

“Deepfake Queen: 2020” Video

https://www.youtube.com/watch?v=IvY-Abd2FfM

“This is not Morgan Freeman” Video

https://www.youtube.com/watch?v=oxXpB9pSETo

Making A Deepfake With DeepFaceLab

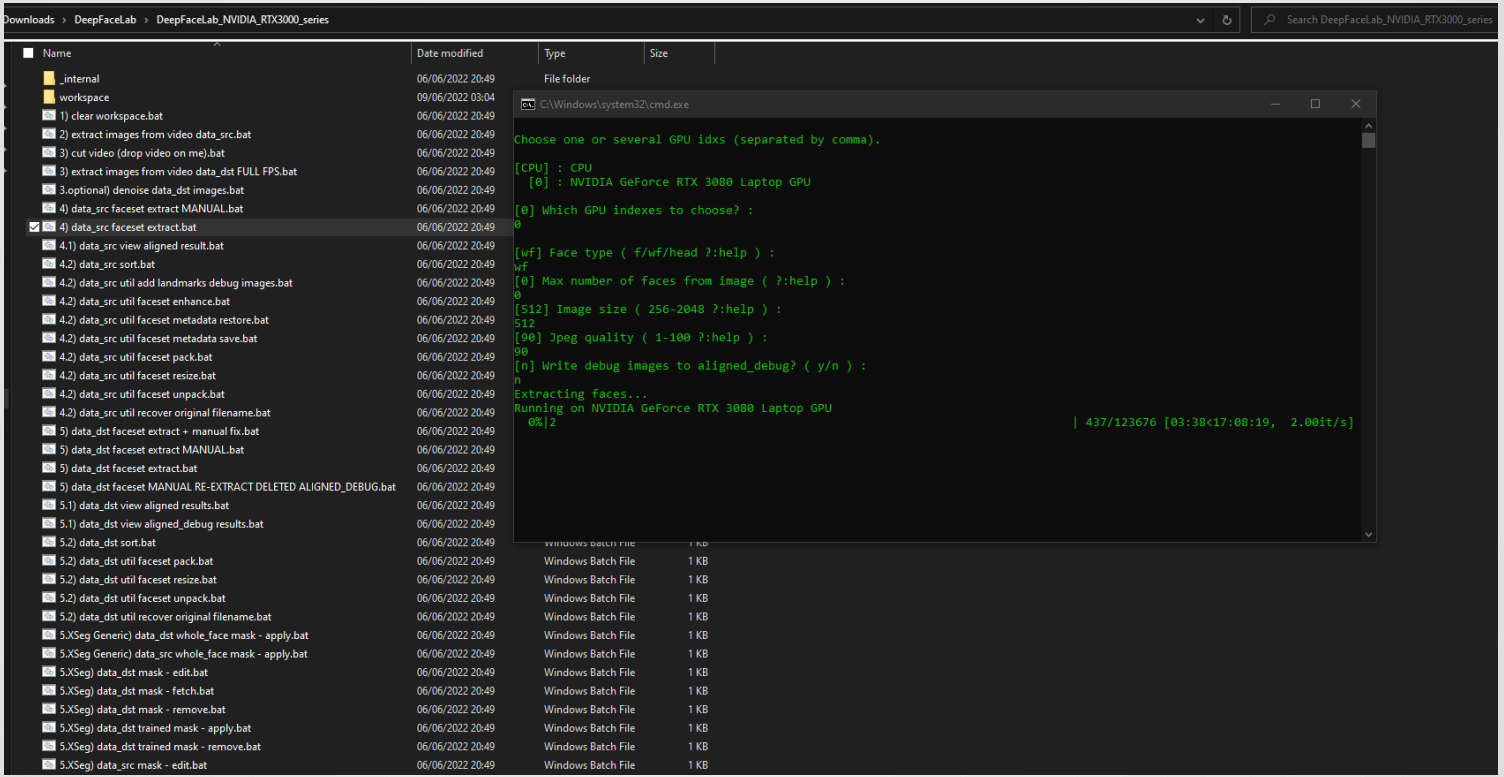
DeepFaceLab is the leading software for creating deepfakes.

https://github.com/iperov/DeepFaceLab

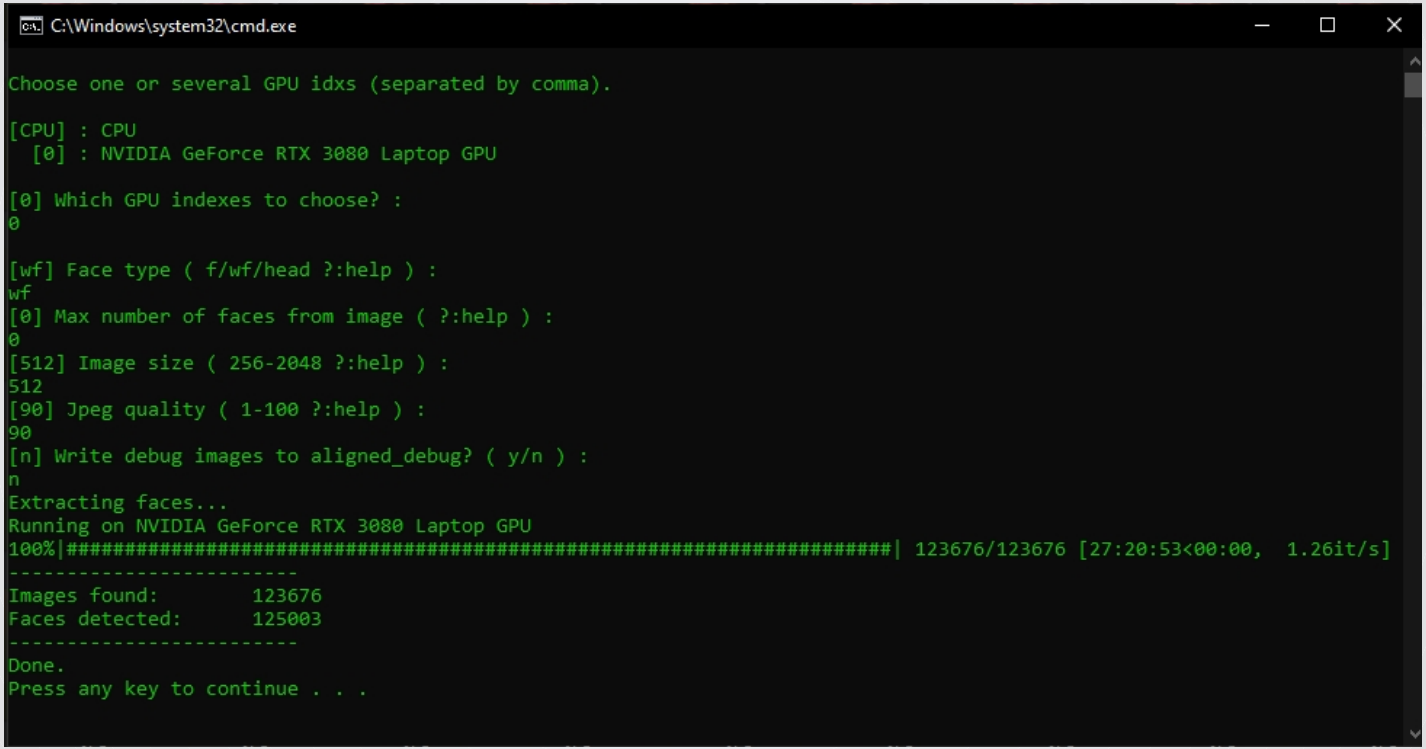
Data Set


Downloads › DeepFaceLab › DeepFaceLab_NVIDIA_RTX3000_series › workspace › data_src
Search data_src
aligned
00001.png 00002.png 00003.png 00004.png 00005.png 00006.png 00007.png 00008.png 00009.png 00010.png 00011.png 00012.png 00013.png 00014.png 00015.png 00016.png 00017.png 00018.png 00019.png 00020.png 00021.png
00022.png 00023.png 00024.png 00025.png 00026.png 00027.png 00028.png 00029.png 00030.png 00031.png 00032.png 00033.png 00034.png 00035.png 00036.png 00037.png 00038.png 00039.png 00040.png 00041.png 00042.png 00043.png
00044.png 00045.png 00046.png 00047.png 00048.png 00049.png 00050.png 00051.png 00052.png 00053.png 00054.png 00055.png 00056.png 00057.png 00058.png 00059.png 00060.png 00061.png 00062.png 00063.png 00064.png 00065.png
00066.png 00067.png 00068.png 00069.png 00070.png 00071.png 00072.png 00073.png 00074.png 00075.png 00076.png 00077.png 00078.png 00079.png 00080.png 00081.png 00082.png 00083.png 00084.png 00085.png 00086.png 00087.png
00088.png 00089.png 00090.png 00091.png 00092.png 00093.png 00094.png 00095.png 00096.png 00097.png 00098.png 00099.png 00100.png 00101.png 00102.png 00103.png 00104.png 00105.png 00106.png 00107.png 00108.png 00109.png
00110.png 00111.png 00112.png 00113.png 00114.png 00115.png 00116.png 00117.png 00118.png 00119.png 00120.png 00121.png 00122.png 00123.png 00124.png 00125.png 00126.png 00127.png 00128.png 00129.png 00130.png 00131.png
00132.png 00133.png 00134.png 00135.png 00136.png 00137.png 00138.png 00139.png 00140.png 00141.png 00142.png 00143.png 00144.png 00145.png 00146.png 00147.png 00148.png 00149.png 00150.png 00151.png 00152.png 00153.png
00154.png 00155.png 00156.png 00157.png 00158.png 00159.png 00160.png 00161.png 00162.png 00163.png 00164.png 00165.png 00166.png 00167.png 00168.png 00169.png 00170.png 00171.png 00172.png 00173.png 00174.png 00175.png
00176.png 00177.png 00178.png 00179.png 00180.png 00181.png 00182.png 00183.png 00184.png 00185.png 00186.png 00187.png 00188.png 00189.png 00190.png 00191.png 00192.png 00193.png 00194.png 00195.png 00196.png 00197.png
00198.png 00199.png 00200.png 00201.png 00202.png 00203.png 00204.png 00205.png 00206.png 00207.png 00208.png 00209.png 00210.png 00211.png 00212.png 00213.png 00214.png 00215.png 00216.png 00217.png 00218.png 00219.png
00220.png 00221.png 00222.png 00223.png 00224.png 00225.png 00226.png 00227.png 00228.png 00229.png 00230.png 00231.png 00232.png 00233.png 00234.png 00235.png 00236.png 00237.png 00238.png 00239.png 00240.png 00241.png
00242.png 00243.png 00244.png 00245.png 00246.png 00247.png 00248.png 00249.png 00250.png 00251.png 00252.png 00253.png 00254.png 00255.png 00256.png 00257.png 00258.png 00259.png 00260.png 00261.png 00262.png 00263.png
00264.png 00265.png 00266.png 00267.png 00268.png 00269.png 00270.png 00271.png 00272.png 00273.png 00274.png 00275.png 00276.png 00277.png 00278.png 00279.png 00280.png 00281.png 00282.png 00283.png 00284.png 00285.png
00286.png 00287.png 00288.png 00289.png 00290.png 00291.png 00292.png 00293.png 00294.png 00295.png 00296.png 00297.png 00298.png 00299.png 00300.png 00301.png 00302.png 00303.png 00304.png 00305.png 00306.png 00307.png
00308.png 00309.png 00310.png 00311.png 00312.png 00313.png 00314.png 00315.png 00316.png 00317.png 00318.png 00319.png 00320.png 00321.png 00322.png 00323.png 00324.png 00325.png 00326.png 00327.png 00328.png 00329.png
00330.png 00331.png 00332.png 00333.png 00334.png 00335.png 00336.png 00337.png 00338.png 00339.png 00340.png 00341.png 00342.png 00343.png 00344.png 00345.png 00346.png 00347.png 00348.png 00349.png 00350.png 00351.png

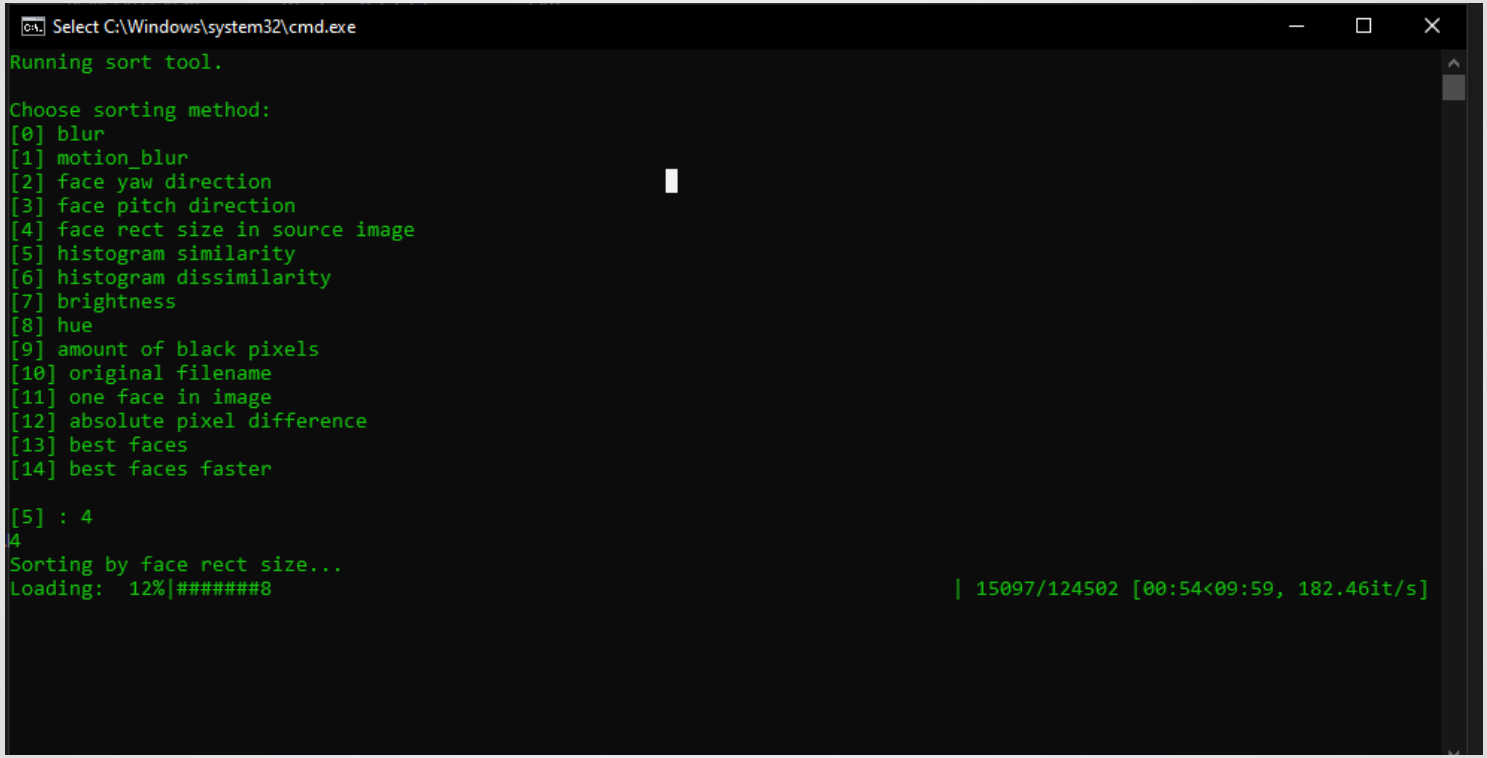
DeepFaceLab | Extract Faces from SRC



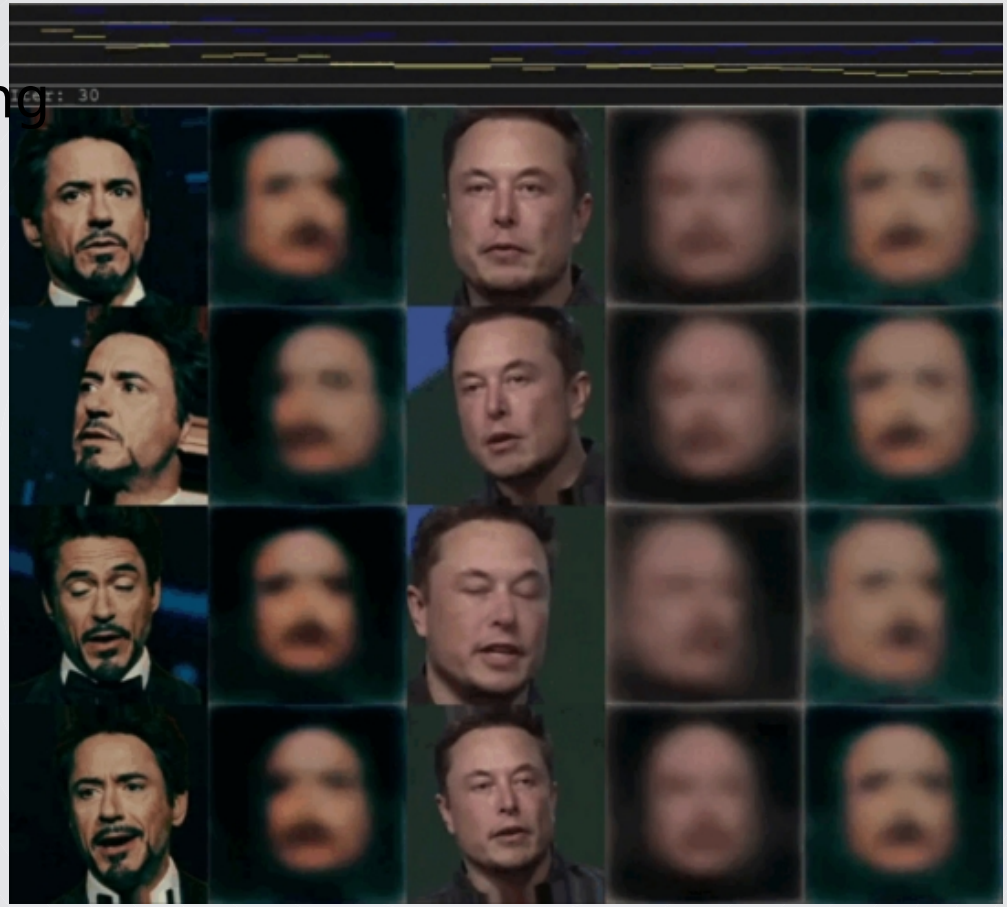
DeepFaceLab | Extract Faces... 27 hours later...



DeepFaceLab | Sorting Faces



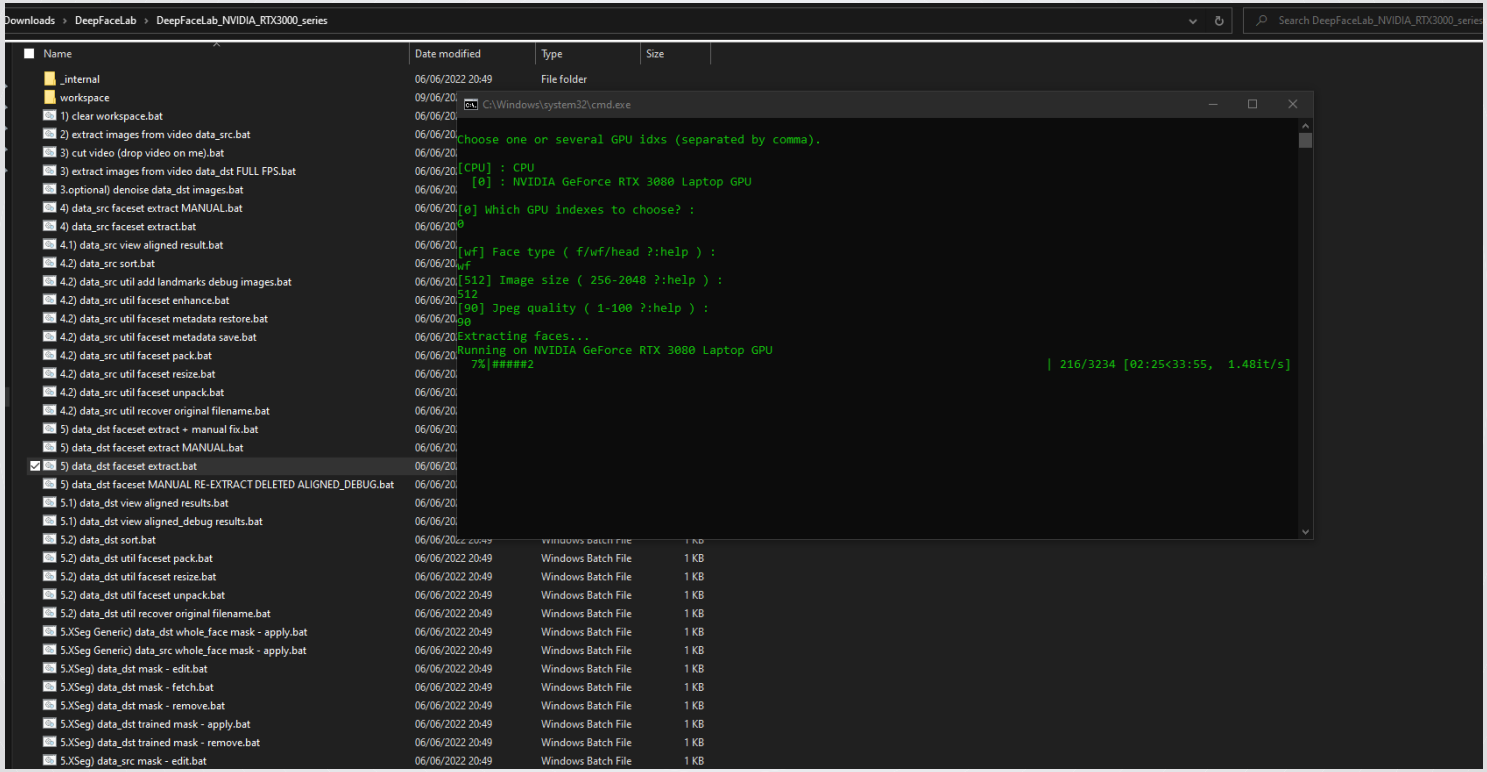
Model Training



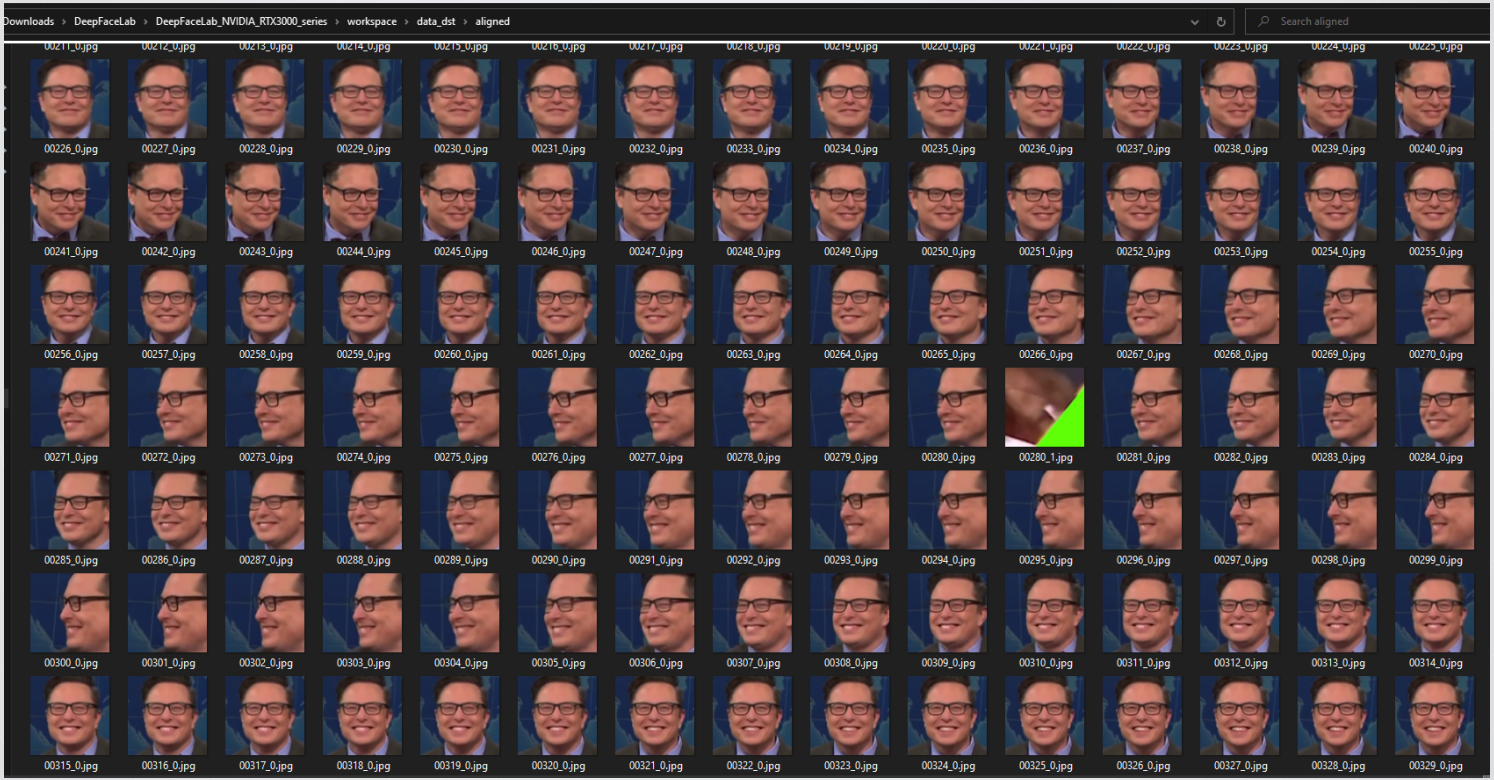
SRC Real SRC Gen DST Real DST Gen SRC + DST Mix

DeepFaceLab | Preprocessing Target Video

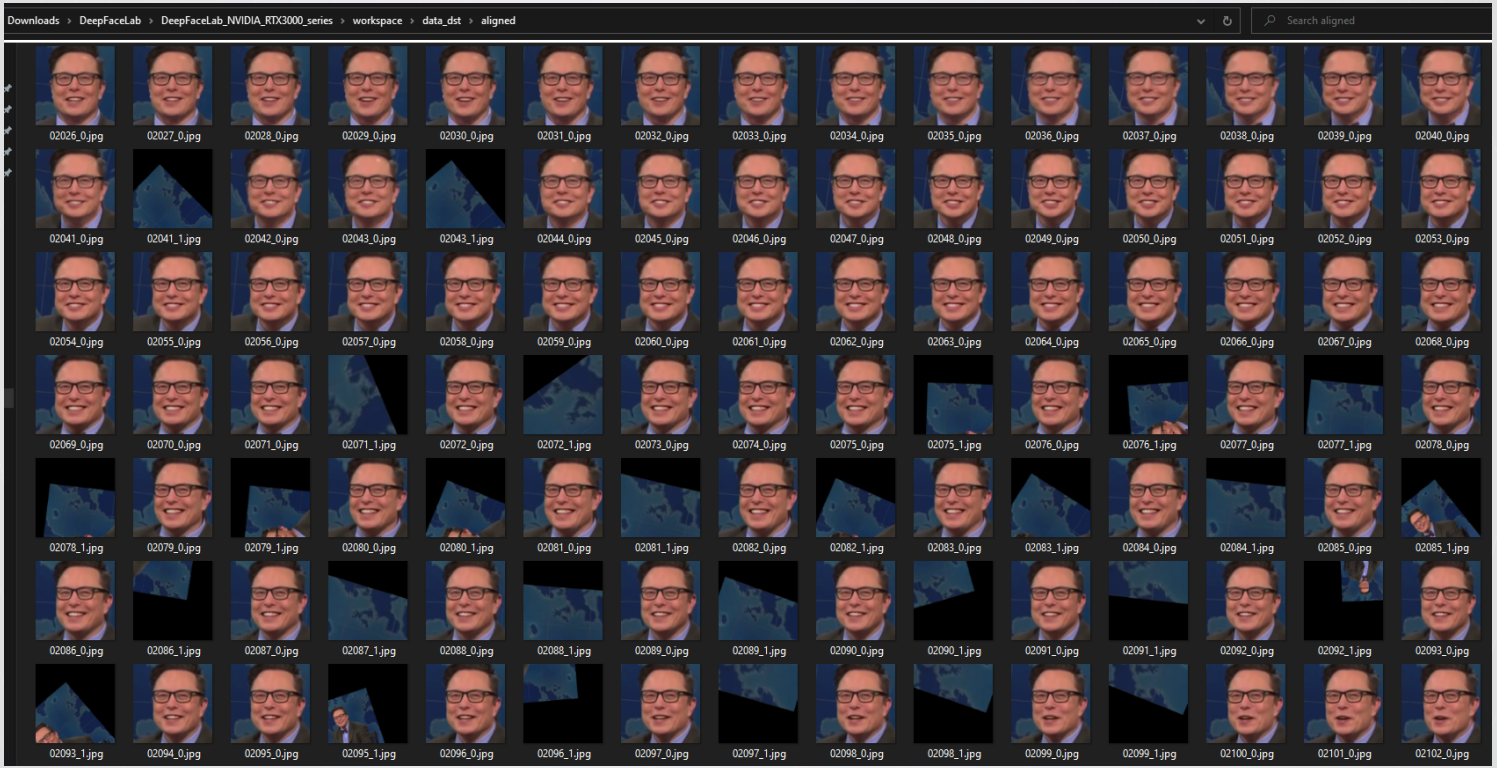
DeepFaceLab | Extracting Faces from DEST



DeepFaceLab | Preprocessing DEST



DeepFaceLab | Preprocessing DEST



Detection

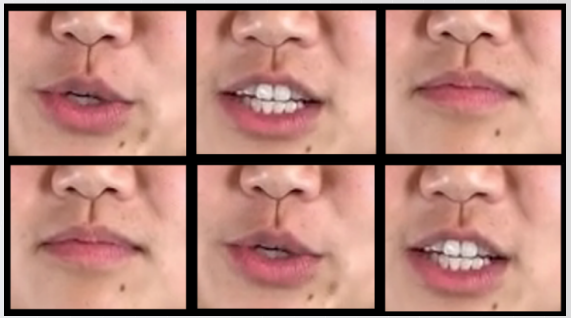
Deepfake Detection

The basic idea is to look for inconsistencies between “visemes,” or mouth formations, and “phonemes,” the phonetic sounds.

Specifically, the researchers looked at the person’s mouth when making the sounds of a “B,” “M,” or “P,” because it’s almost impossible to make those sounds without firmly closing the lips.

Researchers at Stanford have said that their approach is merely part of a “cat-and-mouse” game. As deep-fake techniques improve, they will leave even fewer clues behind.

In the long run, the real challenge is less about fighting deep-fake videos than about fighting disinformation. To reduce disinformation, we need to increase media literacy and develop systems of accountability. For example laws against deliberately producing disinformation and consequences for breaking them, as well as mechanisms to repair the harms caused as a result.



Positive Benefits

Benefits of Deepfakes?

Deepfake technology also has positive uses in many industries, including movies, educational media and digital communications, games and entertainment, social media and healthcare, material science, and various business fields, such as fashion and e-commerce.

The film industry can benefit from deepfake technology in multiple ways. For example, it can help in making digital voices for actors who lost theirs due to disease, or for updating film footage instead of reshooting it.

Deepfakes For Good

(31:00, 33:43)

https://www.youtube.com/watch?v=V5aZjsWM2wo

Questions?