DeepFakes

-Choukha Ram

«deep learning» and «fake»

Deepfake is an AI-based technology used to produce or alter video content so that it presents something that didn't, in fact, occur. The word, which applies to both the technologies and the videos created with it, is a portmanteau of ***deep learning*** **and** ***fake***.

Deepfake

Original Video (https://youtu.be/8LPVjHxXvJM)

Deepfake Video (Cardi Smith)

Types of Deepfakes

Face Replacement/Swapping

- Taking an image of someone’s face (the source) and carefully ‘stitching’ it onto that of another person (the target).
- The identity of the target is concealed, with the focus being on the source.

Face re-enactment

- Manipulating the features of a target’s face, including the movement of their mouth, eyebrows, eyes and the tilting of their head.
- Re-enactment does not aim to replace identities but rather to contort a person’s expressions so they appear to be saying something they are not.

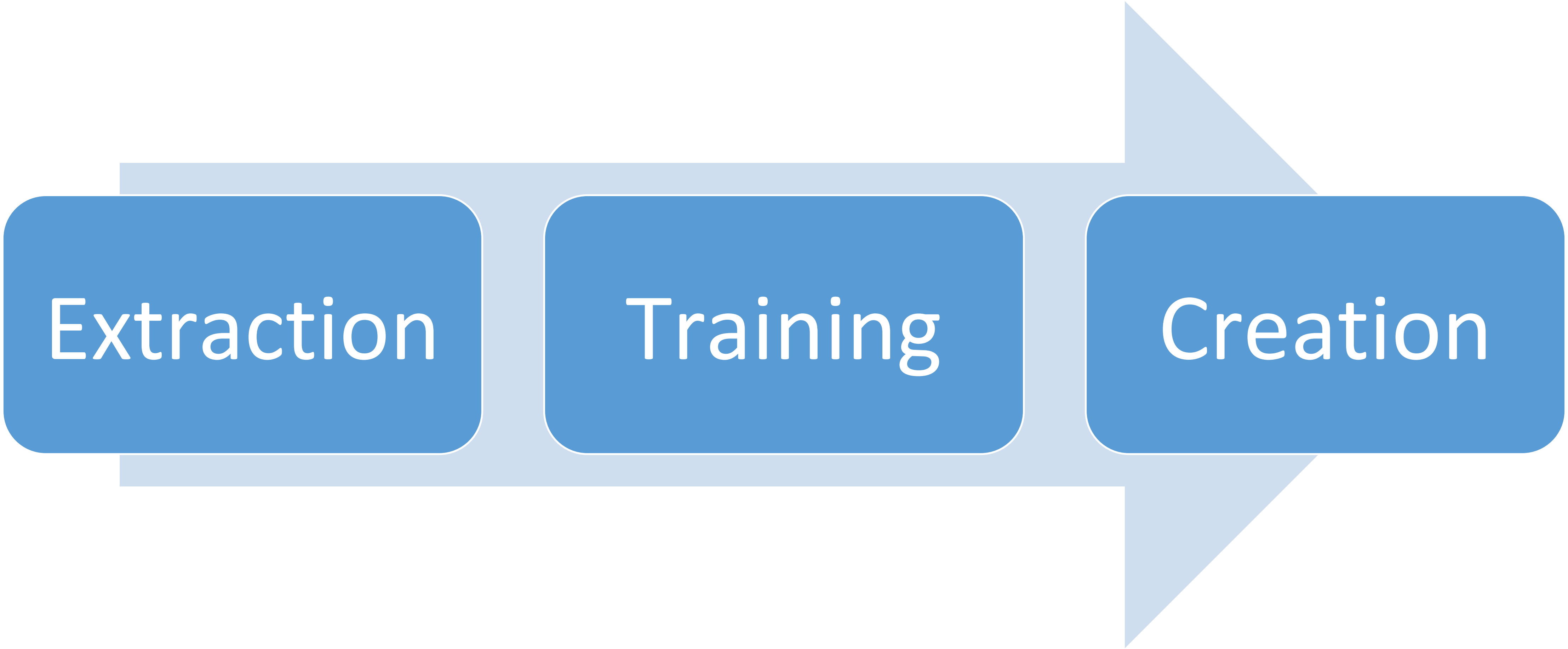
Face generation

- Creating entirely new images of faces using Generative Adversarial Networks.

Speech synthesis

- creating a model of someone’s voice, which can read out text in the same manner, intonation and cadence as the target person.

How deepfakes are created ?



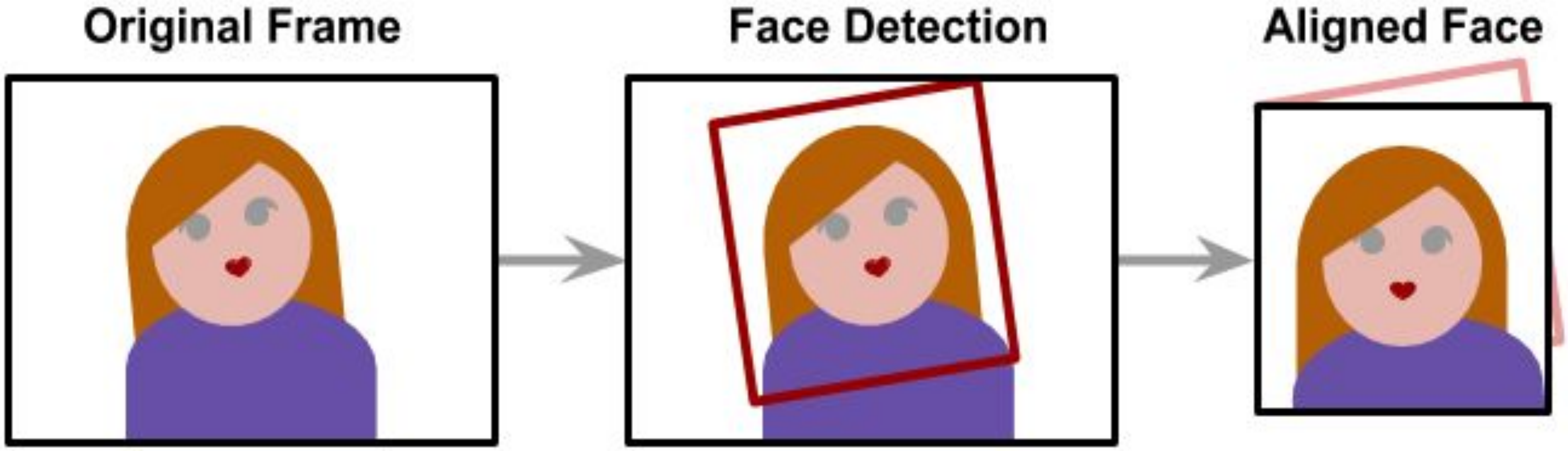
Extraction

Training

Creation

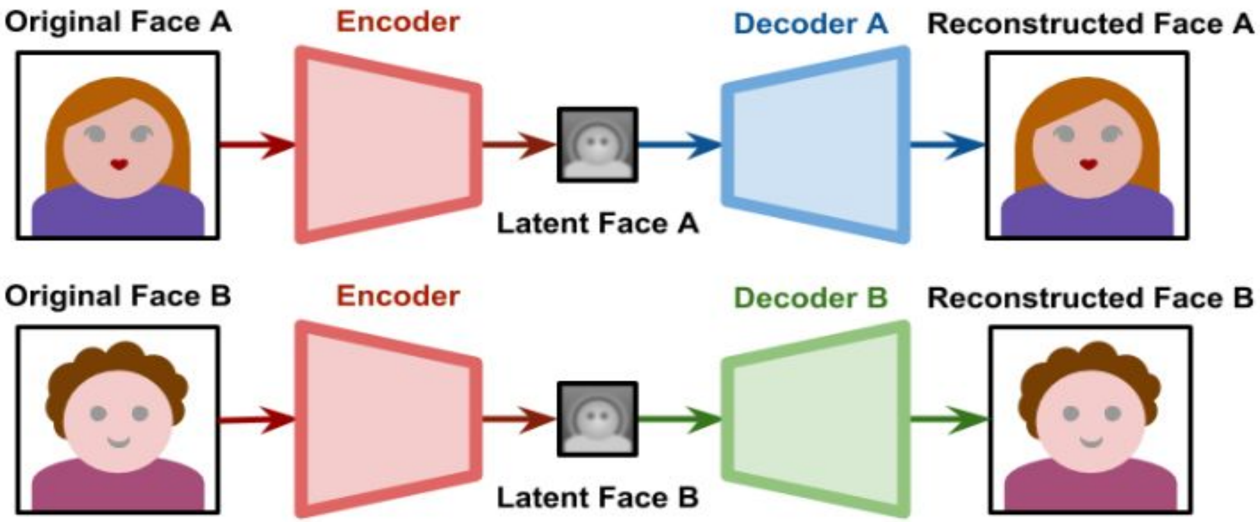
Extraction

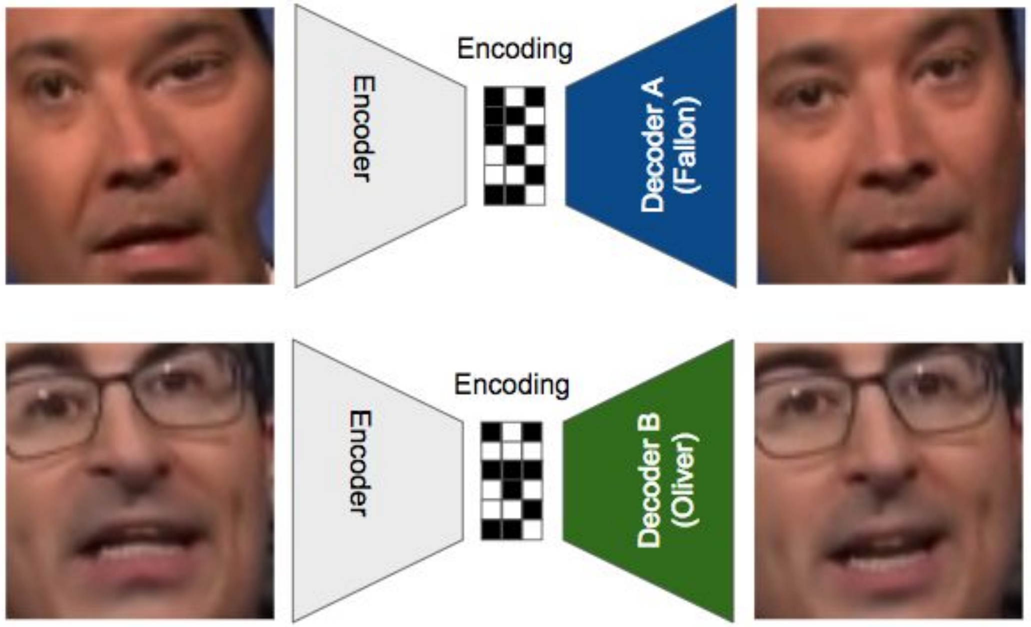
- The first task is to collect sufficient images on which to train the face replacement model. A popular method is to start with videos of both the source and target persons, cut these into individual frames, and then crop the images so only a portrait of the face remains. Ideally the two individuals would resemble each other in face structure and head size.



Training

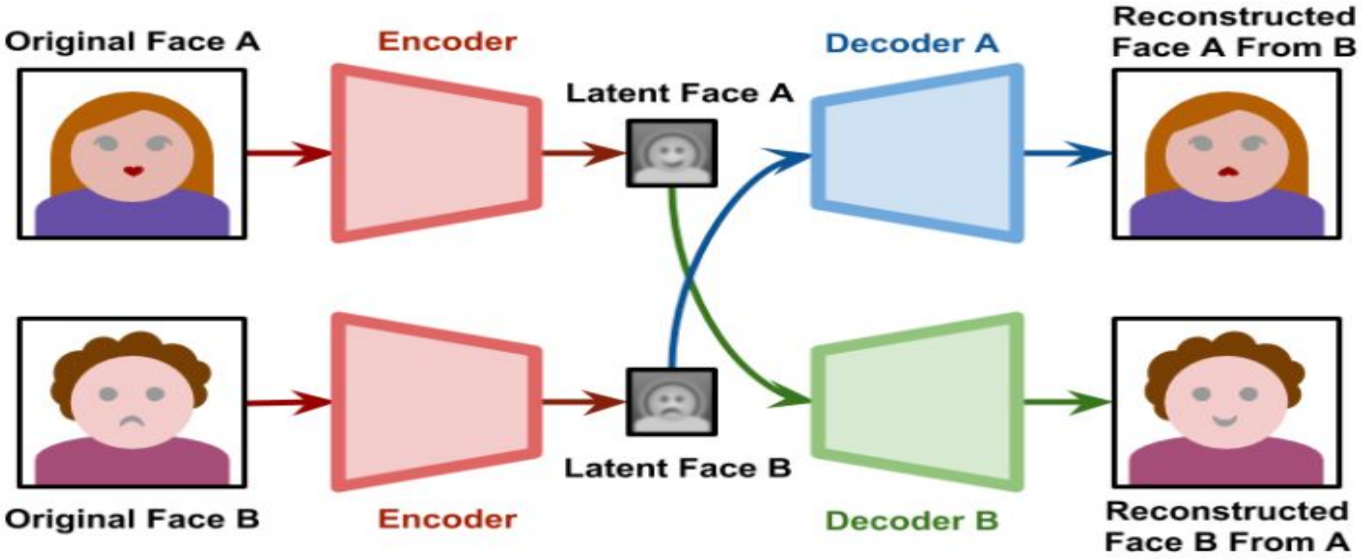
- The next task is to train the face replacement model using the images collected. This is done using 2 **autoencoder** networks, which is a neural network made up of two parts: an **encoder** and a **decoder**.
- During the training phase, these two networks are treated separately. The **Decoder A** is only trained with faces of A; the **Decoder B** is only trained with faces of B. However, all latent faces are produced by the same **Encoder**. This means that the encoder itself has to identify common features in both faces.

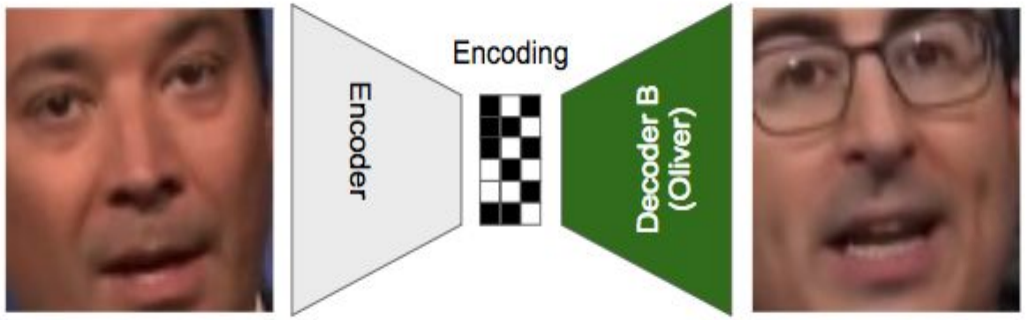




Creation

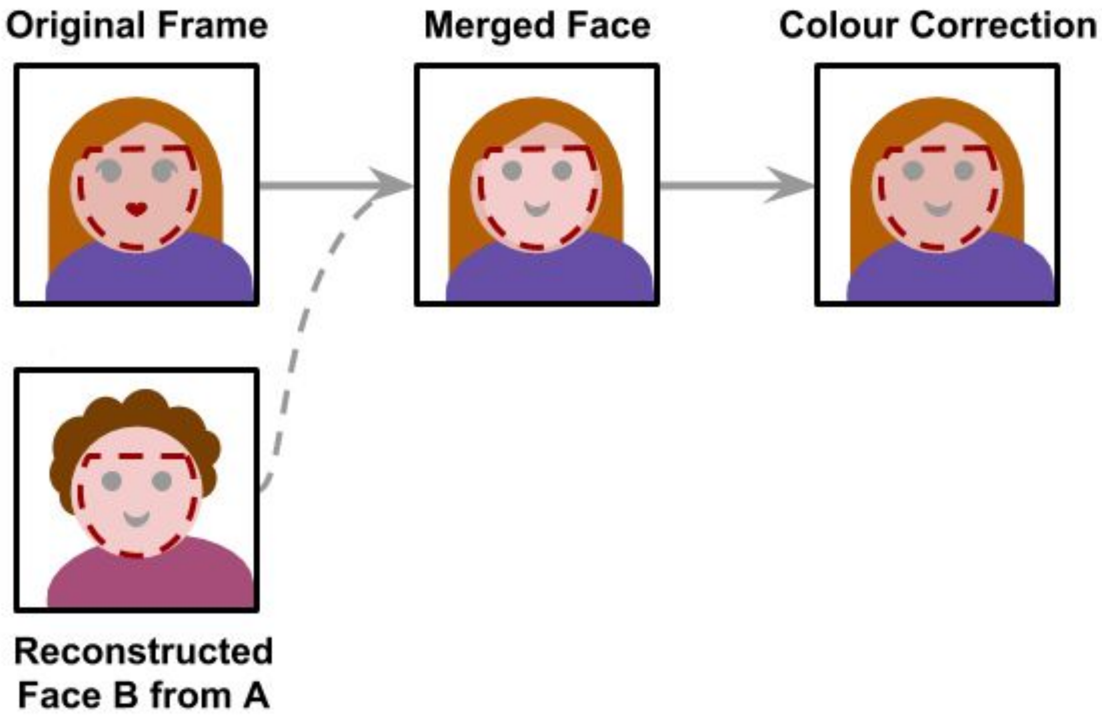
- When the training process is complete, we can pass a latent face generated from Subject A to the Decoder B. As seen in the diagram below, the Decoder B will try to reconstruct Subject B, from the information relative to Subject A.
- If the network has generalised well enough what makes a face, the latent space will represent facial expressions and orientations. This means generating a face for Subject B with the same expression and orientation of Subject A.



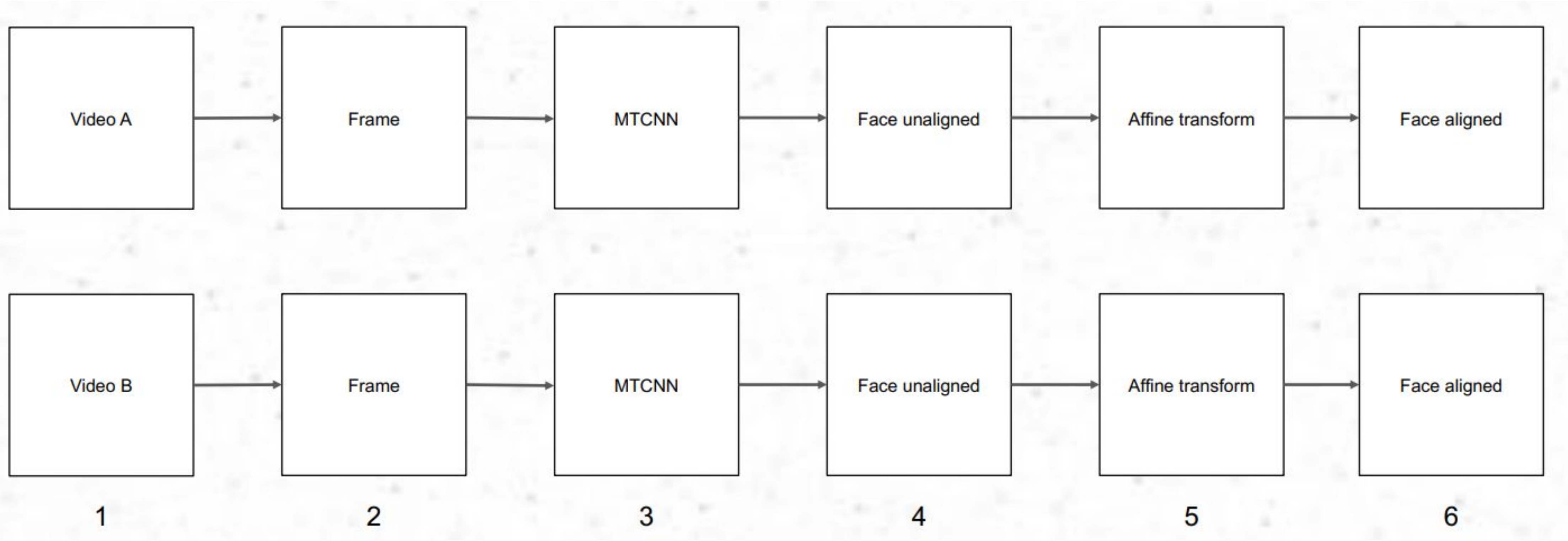


Creation

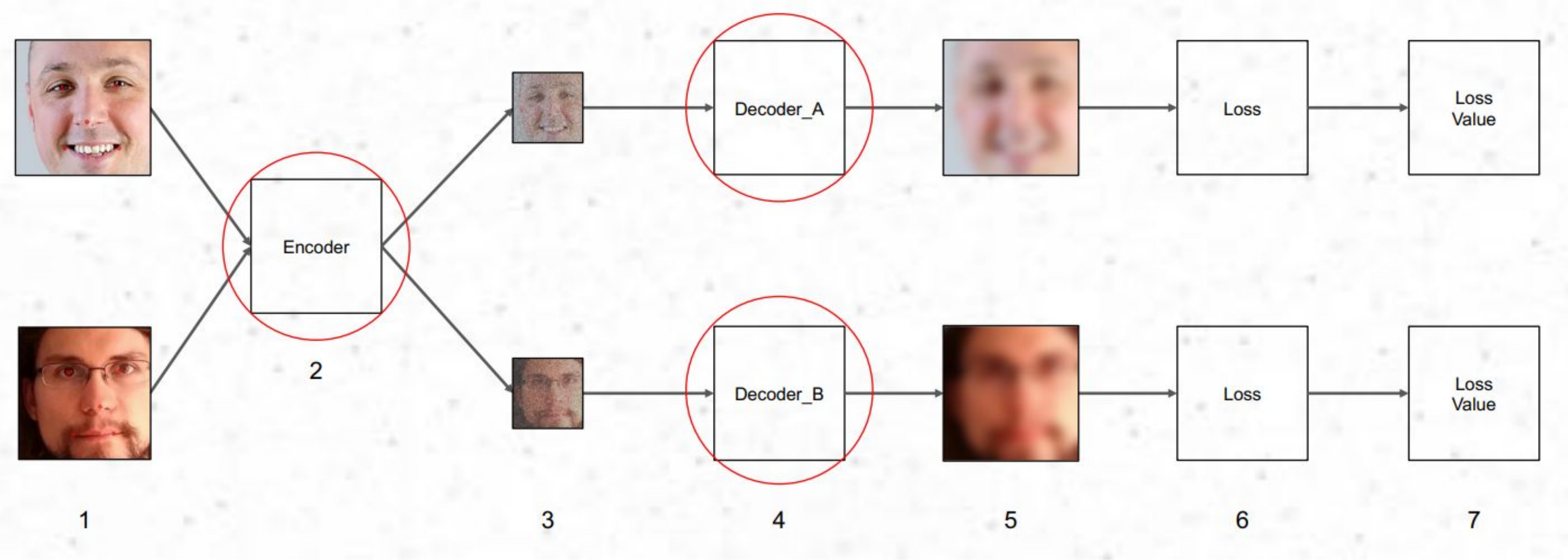
- Starting from a video, all frames are extracted and all faces are aligned. Then, each one is converted using the trained neural network. The final step is to merge the converted face back into the original frame.



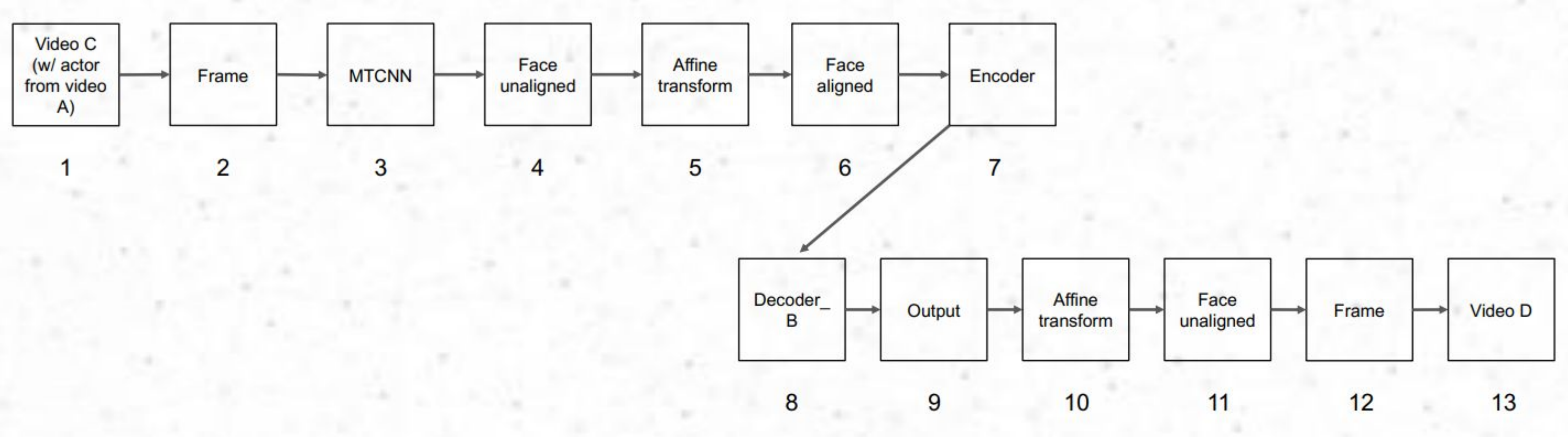
Extraction


Video A
Frame
MTCNN
Face unaligned
Affine transform
Face aligned
Video B
Frame
MTCNN
Face unaligned
Affine transform
Face aligned
1
2
3
4
5
6

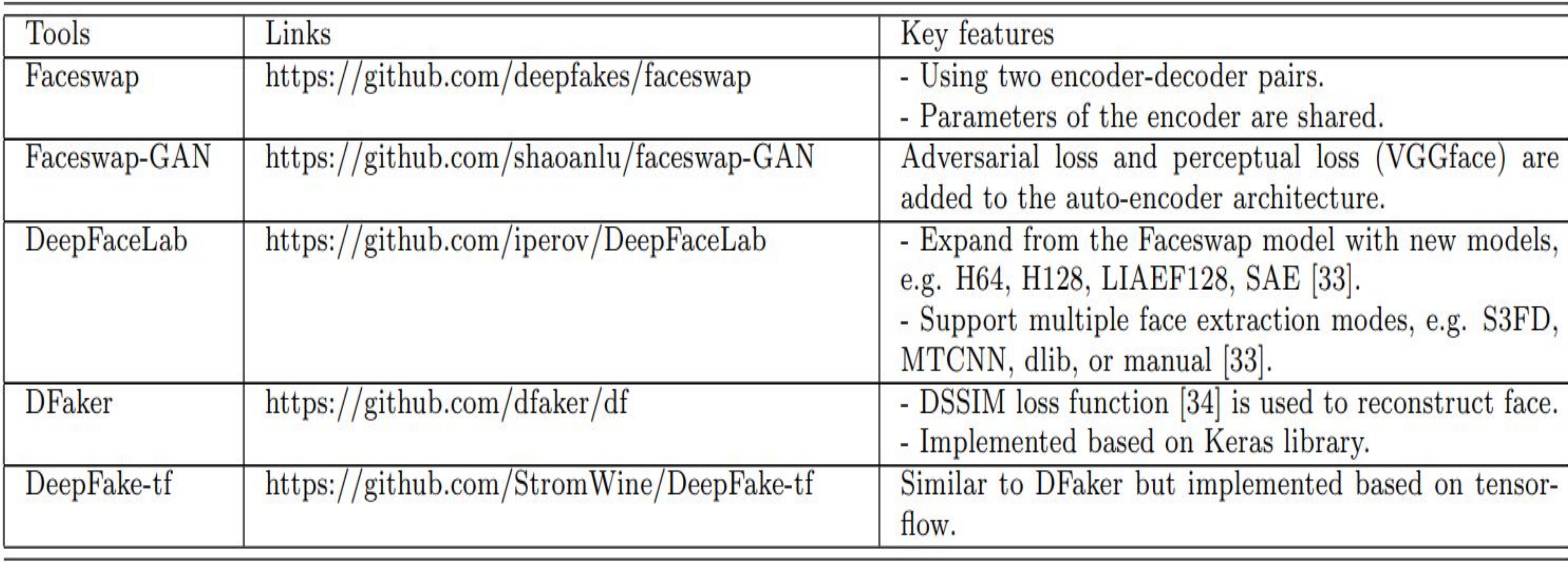
Training


Encoder
Decoder_A
Decoder_B
Loss
Loss Value
Loss
Loss Value
1
2
3
4
5
6
7

Conversion


Video C (w/ actor from video A)
1
Frame
2
MTCNN
3
Face unaligned
4
Affine transform
5
Face aligned
6
Encoder
7
Decoder_ B
8
Output
9
Affine transform
10
Face unaligned
11
Frame
12
Video D
13

Deepfake creation tools



Tools	Links	Key features
Faceswap	https://github.com/deepfakes/faceswap	- Using two encoder-decoder pairs. - Parameters of the encoder are shared.
Faceswap-GAN	https://github.com/shaoanlu/faceswap-GAN	Adversarial loss and perceptual loss (VGGface) are added to the auto-encoder architecture.
DeepFaceLab	https://github.com/iperov/DeepFaceLab	- Expand from the Faceswap model with new models, e.g. H64, H128, LIAEF128, SAE [33]. - Support multiple face extraction modes, e.g. S3FD, MTCNN, dlib, or manual [33].
DFaker	https://github.com/dfaker/df	- DSSIM loss function [34] is used to reconstruct face. - Implemented based on Keras library.
DeepFake-tf	https://github.com/StromWine/DeepFake-tf	Similar to DFaker but implemented based on tensorflow.

What deepfakes can be used for ? (Good)

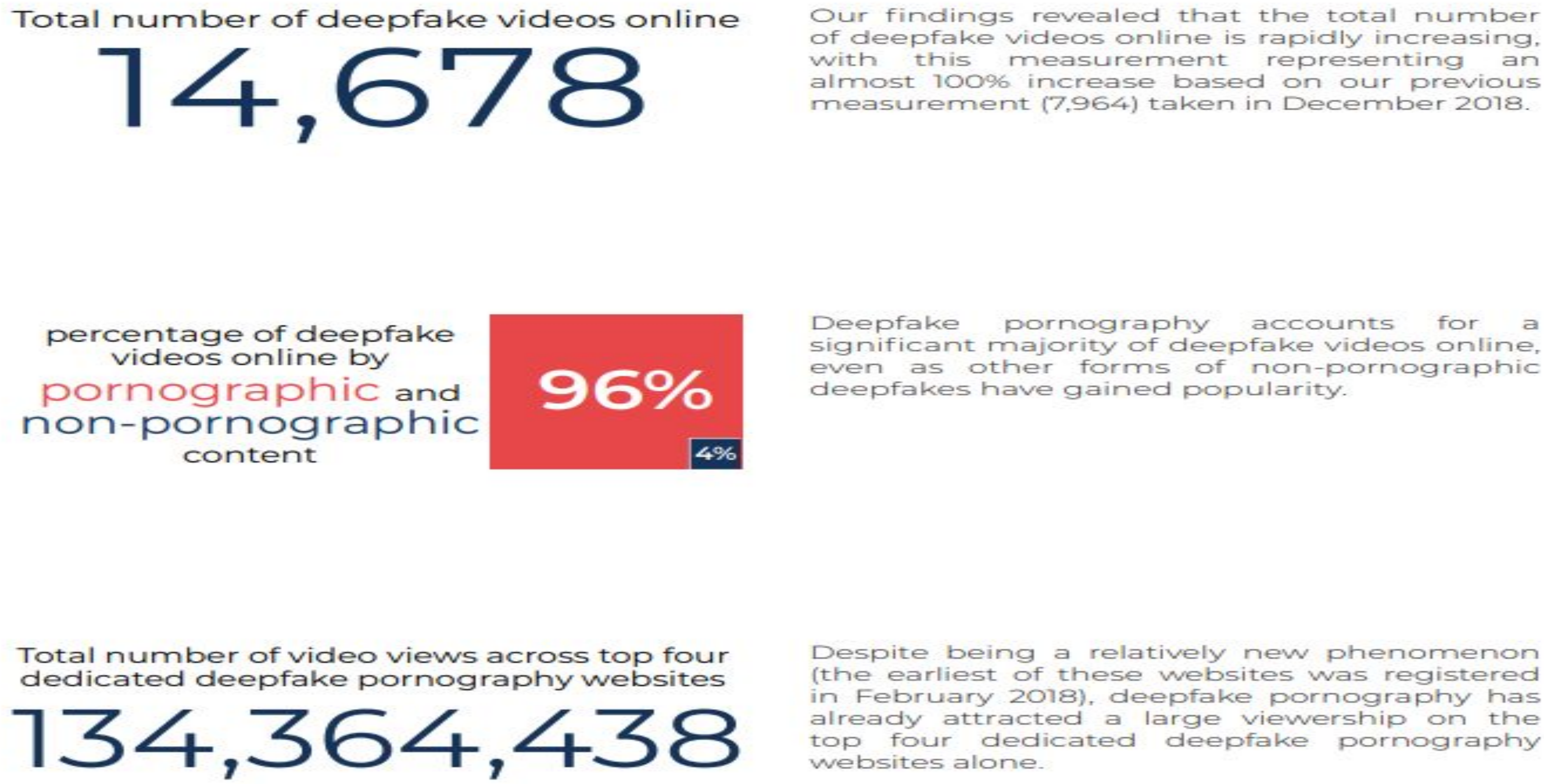
- **Video Content Production** – movies with unknown actors & then replacing with celebrities, custom content based on demand,to bring deceased actors ‘back to life’ for film sequels.
- **Social Apps** – Instagram/Snapchat filters, personalized greetings
- **Licensing Celebrity Faces** - for fashion & modelling work
- **Personalized Advertising** - Imagine a world where the ads you see as you surf the web include you, your friends, and your family.
- **Healthcare** - creating artificial voice replacements that people can adopt once they lose the ability to speak.

Why deepfakes are a problem ?

- **Deepfake Pornography** - Deepfake technology is being weaponized against men/women by inserting their faces into porn.
- **DeepNude** – DeepNude, a computer app that enables users to ‘strip’ photos of clothed women. The app uses deep learning image translation algorithms that have been tuned to synthetically remove clothes from images of women, and generate naked parts of their body that were previously covered.
- **Politics** – manipulating elections & sentiments of people.
- **Crime** - blackmail, Counterfeiting of evidence, real criminals hiding behind deepfakes
- **Cybersecurity:**
  - Enhancing fake digital identities: fraud, infiltration and espionage
  - Synthetic voice impersonation and fraud

State of Deepfakes (Videos)

- Deeptrace report Sep. 2019



Total number of deepfake videos online

14,678

Our findings revealed that the total number of deepfake videos online is rapidly increasing, with this measurement representing an almost 100% increase based on our previous measurement (7,964) taken in December 2018.

percentage of deepfake videos online by pornographic and non-pornographic content

96%

4%

Deepfake pornography accounts for a significant majority of deepfake videos online, even as other forms of non-pornographic deepfakes have gained popularity.

Total number of video views across top four dedicated deepfake pornography websites

134,364,438

Despite being a relatively new phenomenon (the earliest of these websites was registered in February 2018), deepfake pornography has already attracted a large viewership on the top four dedicated deepfake pornography websites alone.

What need to be done about deepfakes?

- **Legislation** - Legislators around the world are considering new laws to suppress audio and visual disinformation.
- **Detection** - Media forensic methods have long been used in criminal courts to interrogate visual evidence, but they can also be applied to help identify deepfakes. Or we need strong AI based tools to detect deepfakes.
- **Education** - Although many people will naturally become more attuned to the presence of doctored footage and begin to view online content more critically as a result, some groups may benefit from having the phenomenon of deepfakes brought to their attention.

References

- https://www.alanzucconi.com/2018/03/14/understanding-the-technology-behind-deepfakes/
- https://goberoi.com/exploring-deepfakes-20c9947c22d9
- https://i.blackhat.com/USA-19/Thursday/us-19-Price-Playing-Offense-And-Defense-With-Deepfakes.pdf
- http://cs230.stanford.edu/projects_spring_2019/reports/18681213.pdf
- https://arxiv.org/pdf/1909.11573.pdf
- https://regmedia.co.uk/2019/10/08/deepfake_report.pdf