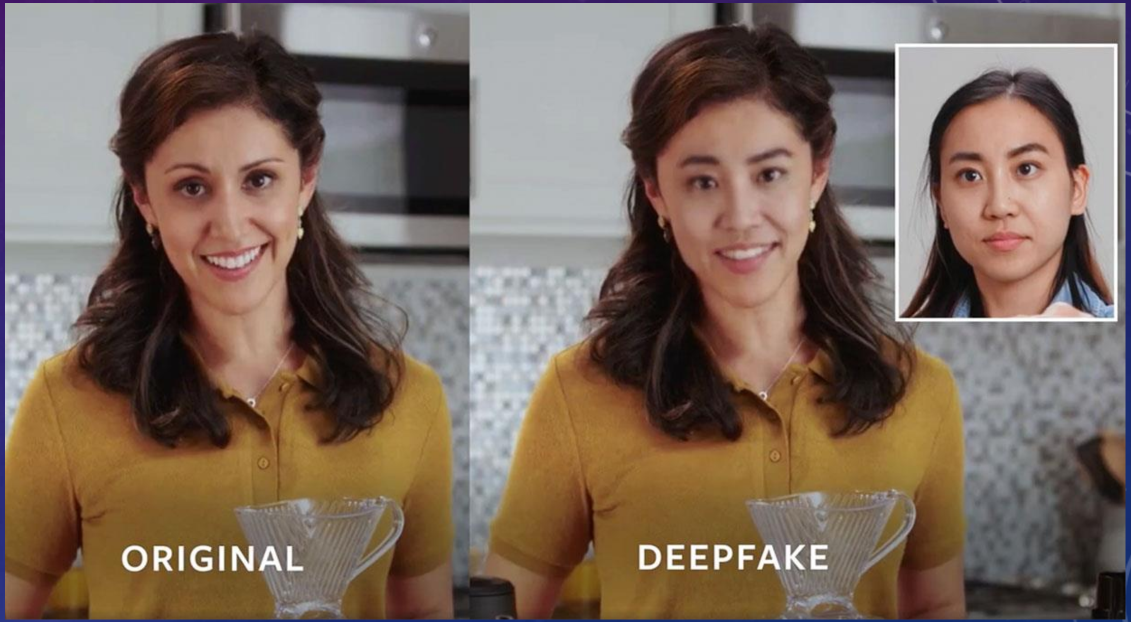
# FACEBOOK DEEPFAKES DETECTION CHALLENGE USING AI AND MACHINE LEARNING

MICHAEL LEONARDI, ASHUTOSH MISAR, SIVAKUMAR G. PILLAI, CHARLES TAPPERT, & AVERY LEIDER

# WHAT IS A DEEPFAKE?

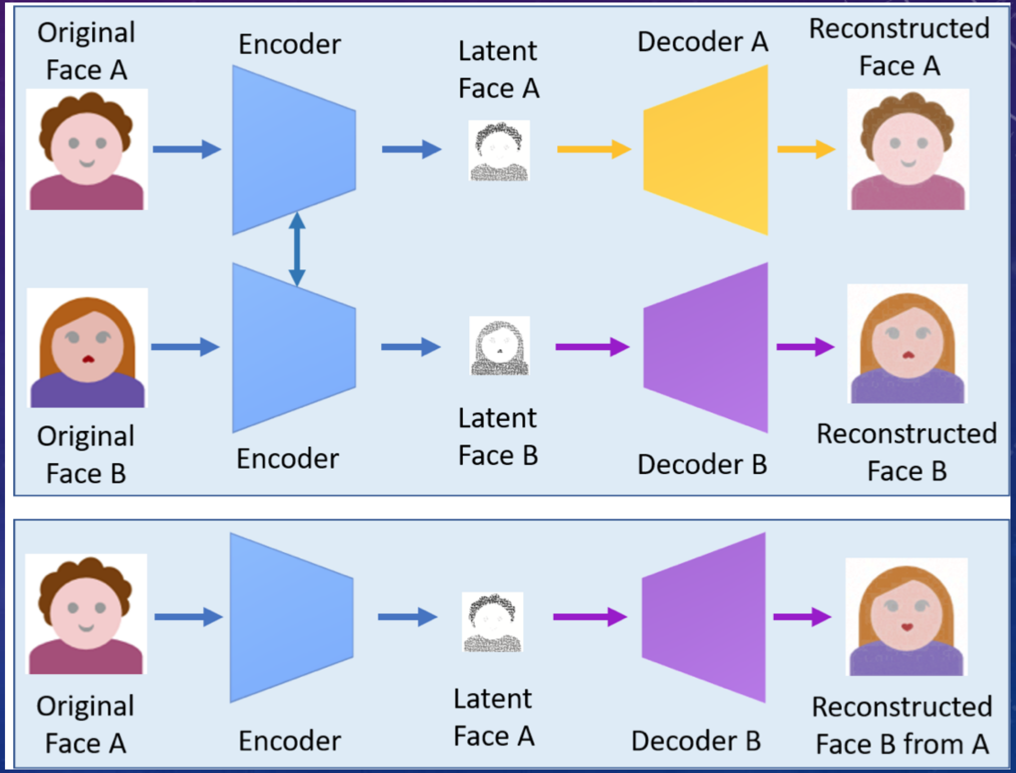
- A Deepfake is an AI-generated fake video which shows someone doing or saying fictitious things.
- This is achieved by changing the face of the actual person in the video and replacing it with a target of your choosing
- A deepfake could be either a  face or voice swap (or both).
- Deepfakes have significant implications for determining the legitimacy of information presented online and yet we don’t have effective methods for detecting fake videos
- The goal of the challenge is to produce technology that everyone can use to better detect when AI has been used to alter a video in order to mislead the viewer



Here is an example of a deep fake image where the face in the original has been replaced with someone else’s face.

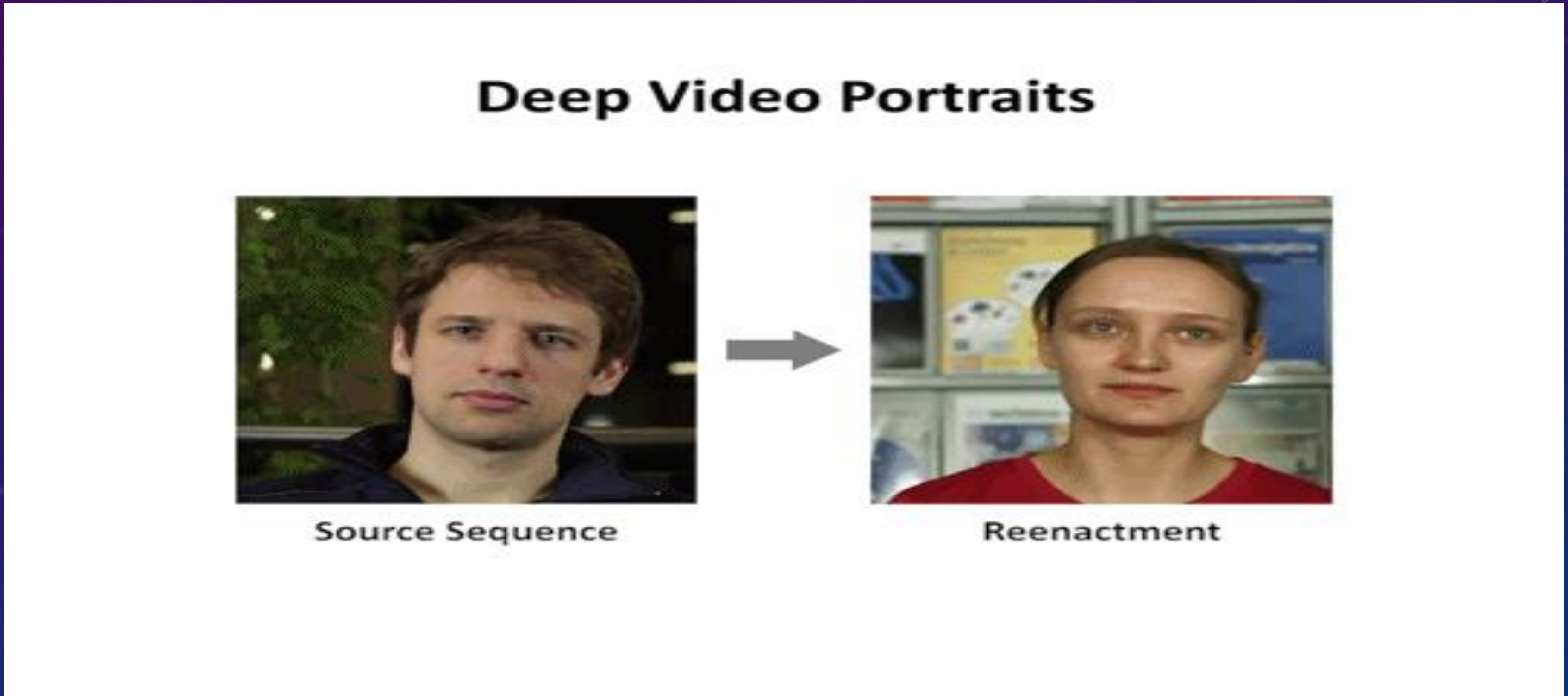
# GENERATIVE ADVERSARIAL NETWORKS (GAN)

- Machine learning systems where two neural networks compete against each other off of a known data set. The larger the dataset, the better fake can be created
- Generator vs Discriminator
  - Generator creates deepfakes to fool discriminator
  - Discriminator attempts to detects video as being fake
  - Process continues until Discriminator cannot recognize the fake video
  - Mimic expressions, blinking and movement of the target
  - The generator has surpassed the discriminator making this a blueprint for how to create deepfakes we cannot detect as being fake



GAN network training process

# EXAMPLE OF ACTUAL GAN OUTPUT



Deep Video Portraits

Source Sequence

Reenactment

Mimic expressions, blinking and movement of the target

# NOTABLE DEEPFAKE VIDEOS

- President Trump lectures Belgium regarding climate change created by a Belgian political party.  This damaged the US relationship with many Belgian politicians until it was shown to be fake
- House speaker Nancy Pelosi giving a press conference appearing to be drunk and slurring words went viral and further agitated the feud between political parties ( President retweeted as mudslinging)
- Facebook creator Mark Zuckerberg being interviewed where he stated that Facebook “owns its users” This was a result of the policy Facebook had created regarding removing fake videos from its sites
- President Obama public service announcement being portrayed by comedian Jordan Peele where he says some outrageous things showing the danger deepfakes present

# OVERALL VIDEO VIEWING STATISTICS VIA FORBES

- One third of all online activity is spent watching videos
- 500 million hours of video are watched on youtube each day
- Half a billion people watch videos on Facebook every day
- 85% of Facebook videos are watched without sound
- Over half of the video content viewed online is done via mobile device
- 92% of mobile video viewers share videos with other people
- The average person looks at a video on social media for less than 10 seconds
- Viewers retain 95% of the message when they watch it in a video

# SOME POTENTIAL DEEPFAKE THREATS

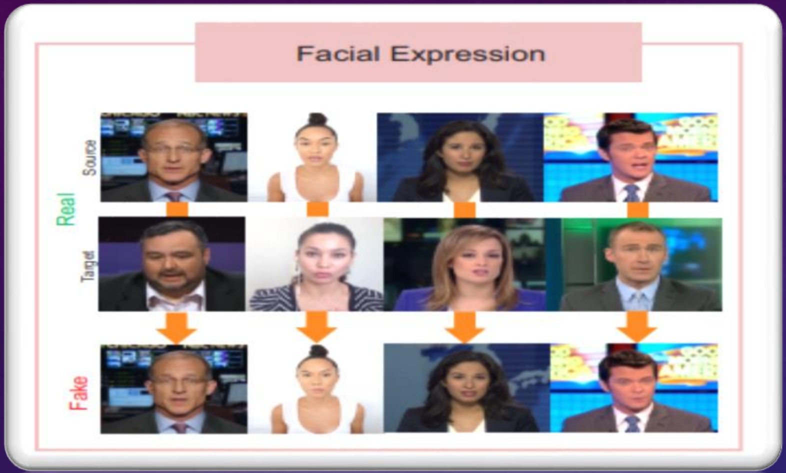
- Increased creation and circulation of fake news
- Spreading inaccurate information quickly via social media
- Negative impact on Presidential Candidates
- Damage to International relations
- False pornographic videos of celebrities (Scarlett Johansson already a victim)
- Portraying infidelity of someone in a marriage
- Showing different groups participating in heinous acts to increase resentment
- Athletes participating in domestic violence or use of illegal substances

# DEEPFAKE CREATION APPLICATIONS

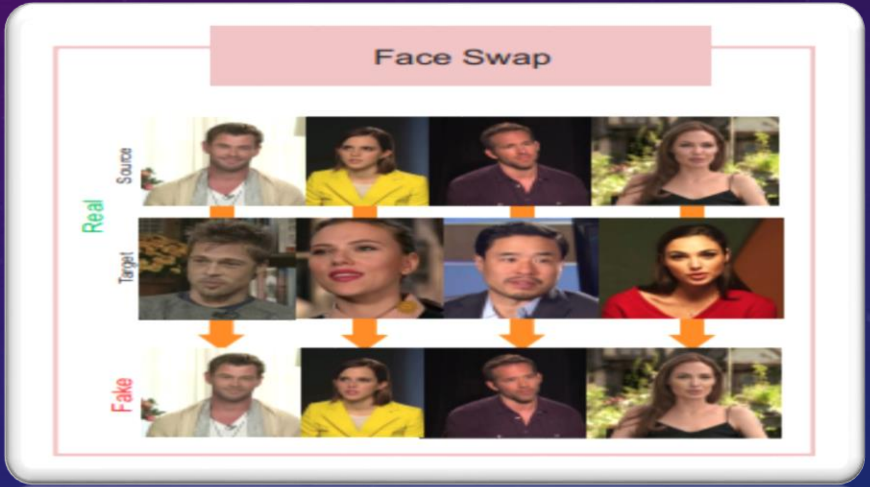
- Zao – Mobile Phone Application to create deep fakes within seconds- only available in China
- Deepfakes Web β – web based service that uses deep learning to absorb the various complexities of face data
- Avenge Them- Swap your face with your favorite Marvel Superhero or Character
- MachineTube- Create Deepfakes from pre-defined models such as Nicholas Cage, Dwayne Johnson, Barack Obama, Kanye West and many others
- DeepFaceLab- Used to train students to use machine learning and image synthesis to replace faces

# FACE MANIPULATION METHODS

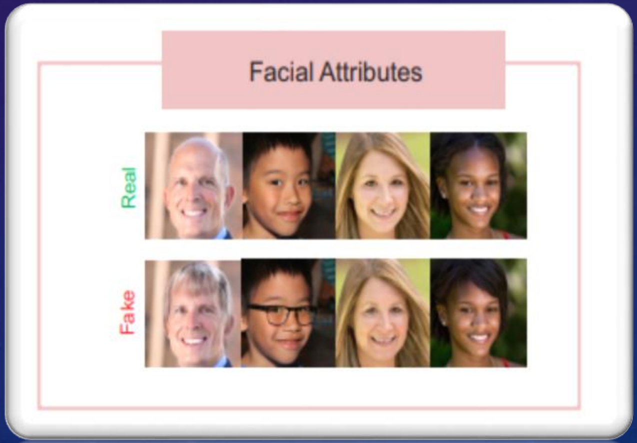
Facial Expression Manipulation ( Face2Face)



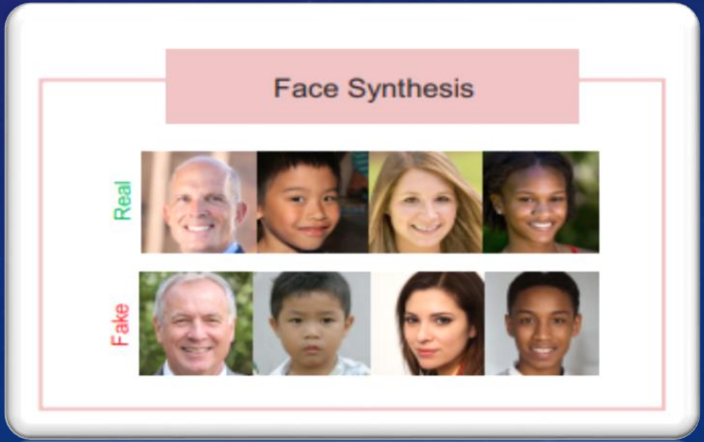
Face Identity Swap(FaceSwap)



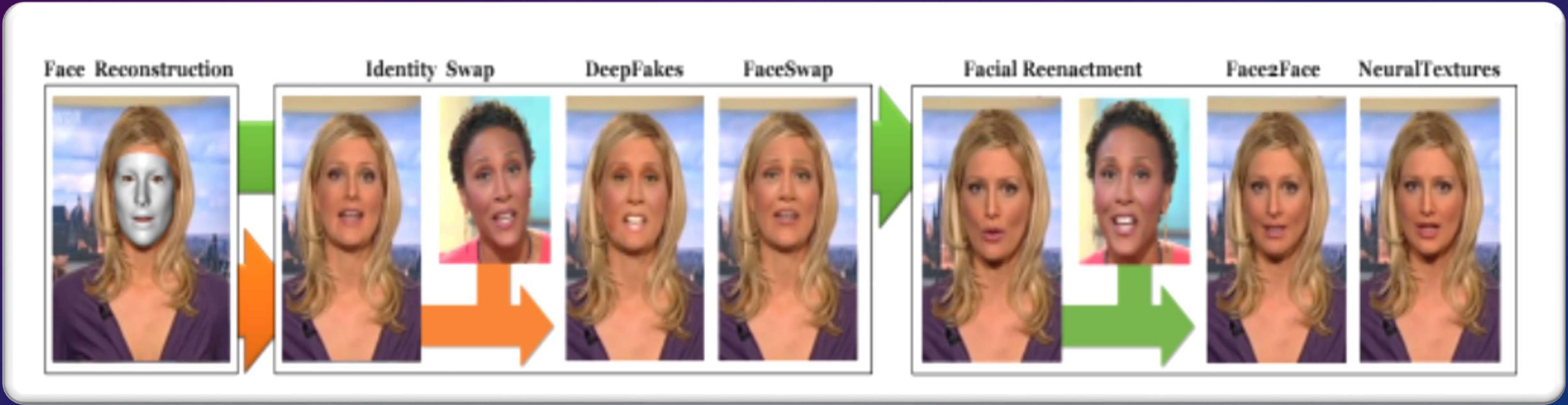
Facial Attributes Manipulation ( Neural Texture)



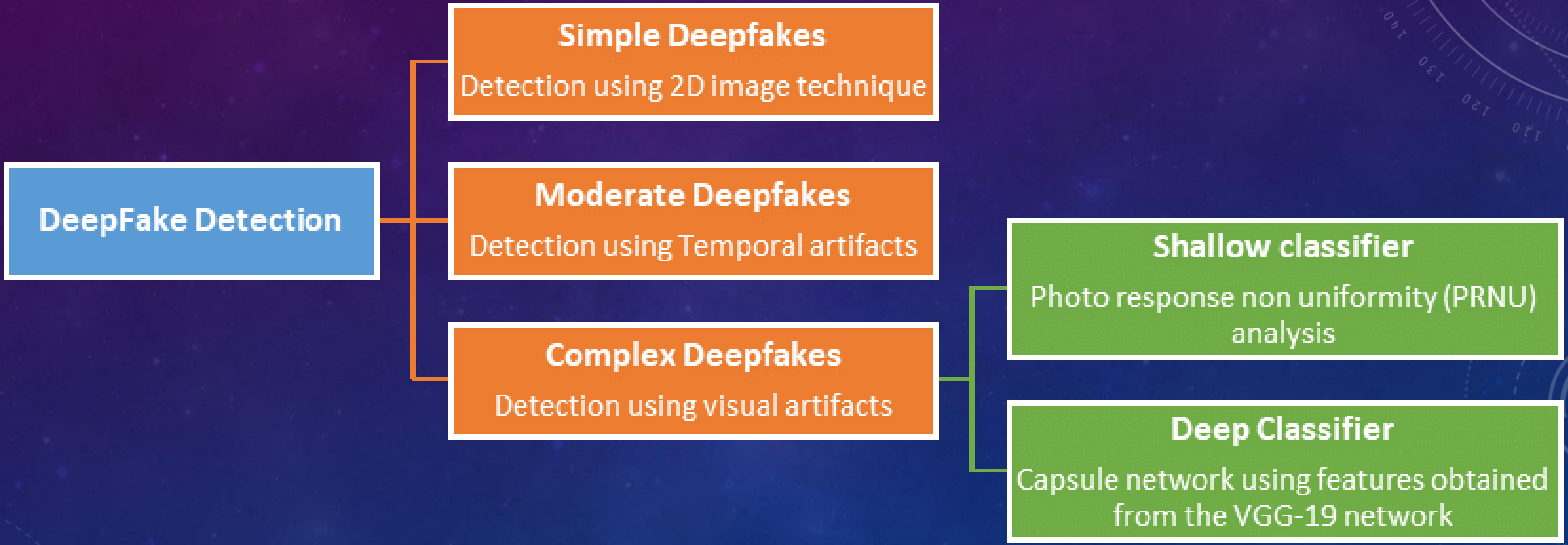
Entire Face Synthesis



# GENERATION OF COMPLEX DEEPFAKES USING FACE MANIPULATION METHODS



# TECHNIQUES TO DETECT DEEPFAKES



DeepFake Detection

**Simple Deepfakes**

Detection using 2D image technique

**Moderate Deepfakes**

Detection using Temporal artifacts

**Complex Deepfakes**

Detection using visual artifacts

**Shallow classifier**

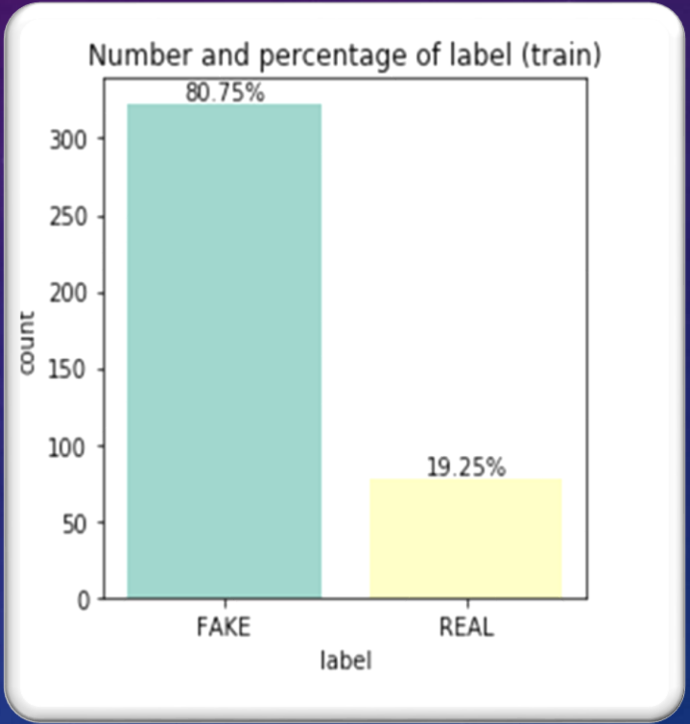
Photo response non uniformity (PRNU) analysis

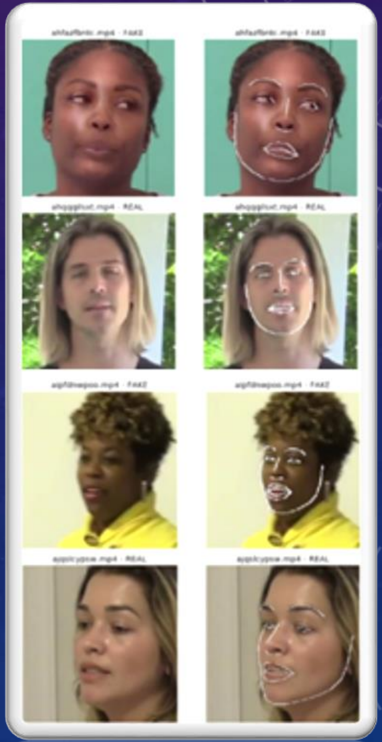
**Deep Classifier**

Capsule network using features obtained from the VGG-19 network

# TECHNIQUES TO DETECT DEEPFAKES

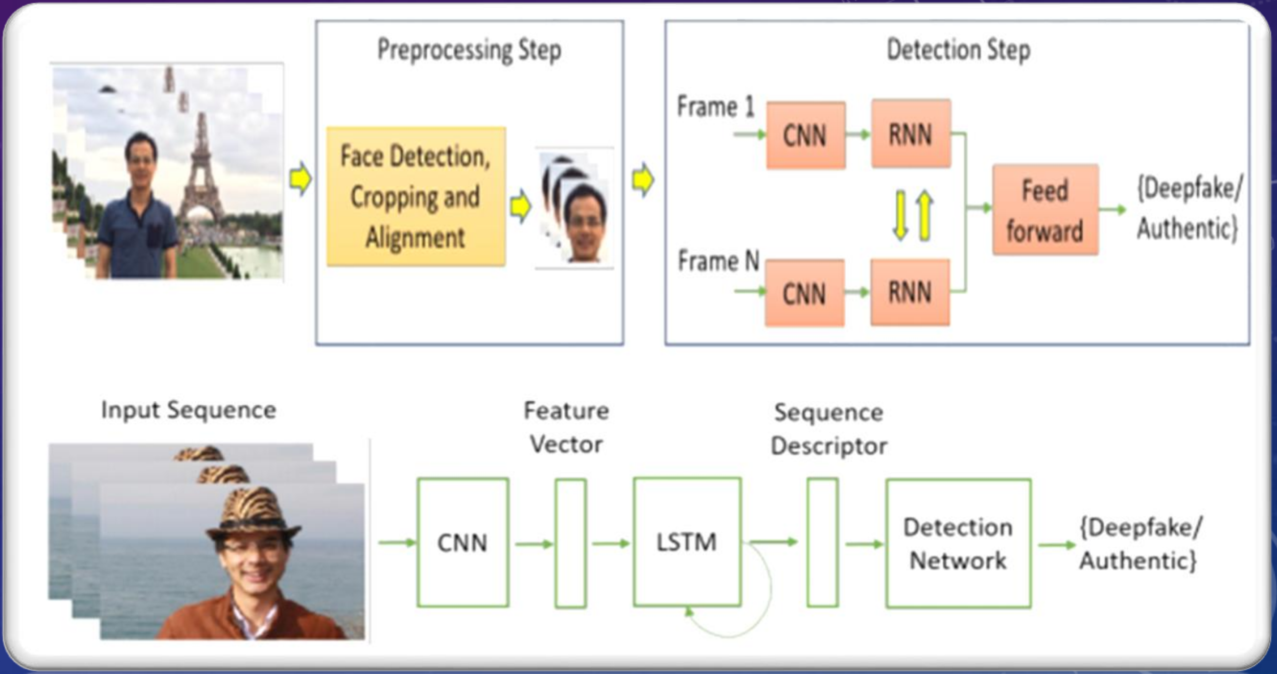
- 2D image face point
  - Takes a single frame from the video and maps the face with key landmarks
  - Repeats the process for all remaining frames in the video
  - Capable of detecting many of the moderate level deepfakes  which account for 30% of fake videos created
  - Not effective in detecting GAN generated fake videos





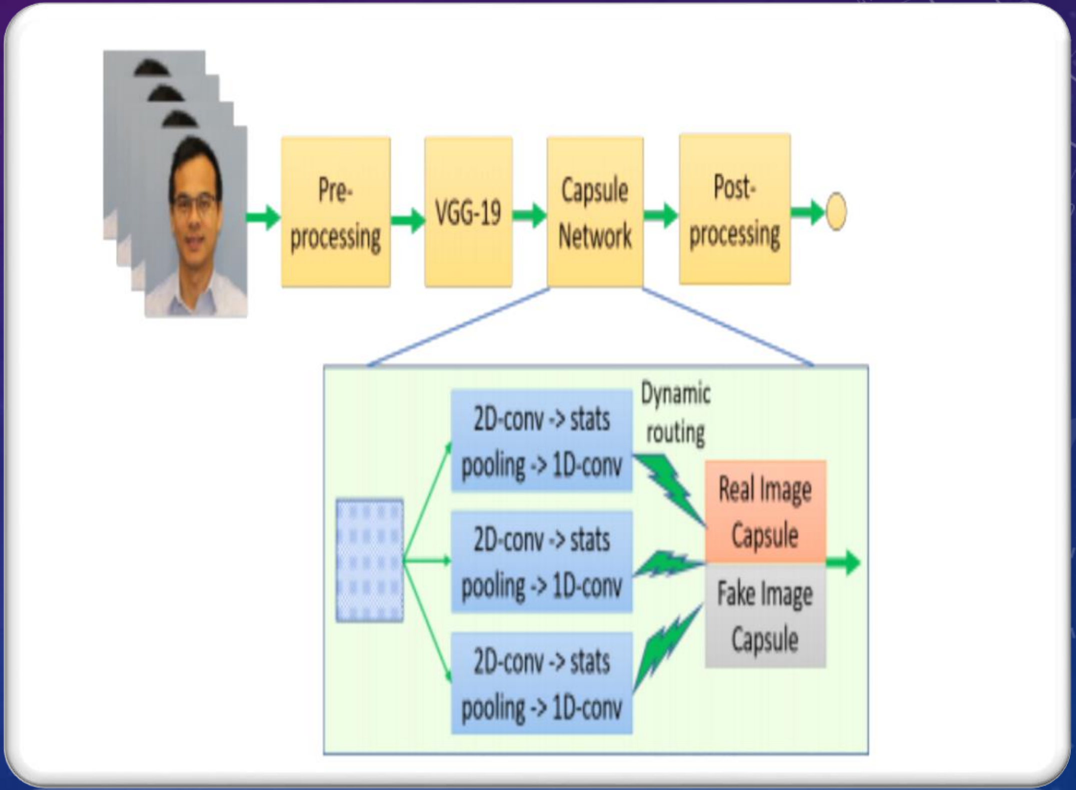
# TECHNIQUES TO DETECT DEEPFAKES

- Detection using Temporal Artifacts (LSTM and CNN)
  - Pre-processing stage in which the objective is to identify, manipulate and align faces on a sequence of frames
  - Distinguishing the manipulated and the authentic facial images with the help of a combination of recurrent neural networks (RNN) along with convolutional neural network(CNN).
  - As there exist temporal inconsistencies and intra-frame inconsistencies between frames of such deep Fake videos, it is appropriate to utilize the temporal-aware pipeline algorithm which implements long short term memory (LSTM) and CNN to identify fake content.



# TECHNIQUES TO DETECT DEEPFAKES

- Detection using visual Artifacts(Capsule Network VGG-19)
  - In order to match the configuration of the authentic videos, Deep Fake videos are generally generated with the limited resolution
  - which is achieved via using affine face wrapping algorithms like scaling, shearing, rotating, etc
  - CNN models like ResNet50, ResNet152, ResNet101, or VGG16 are used to detect the artifacts which are the result of the resolution inconsistency among the wrapped face and the surrounding context
  - In order to differentiate the authentic videos from the fake videos, features of the VGG-19 network are utilized by the capsule network

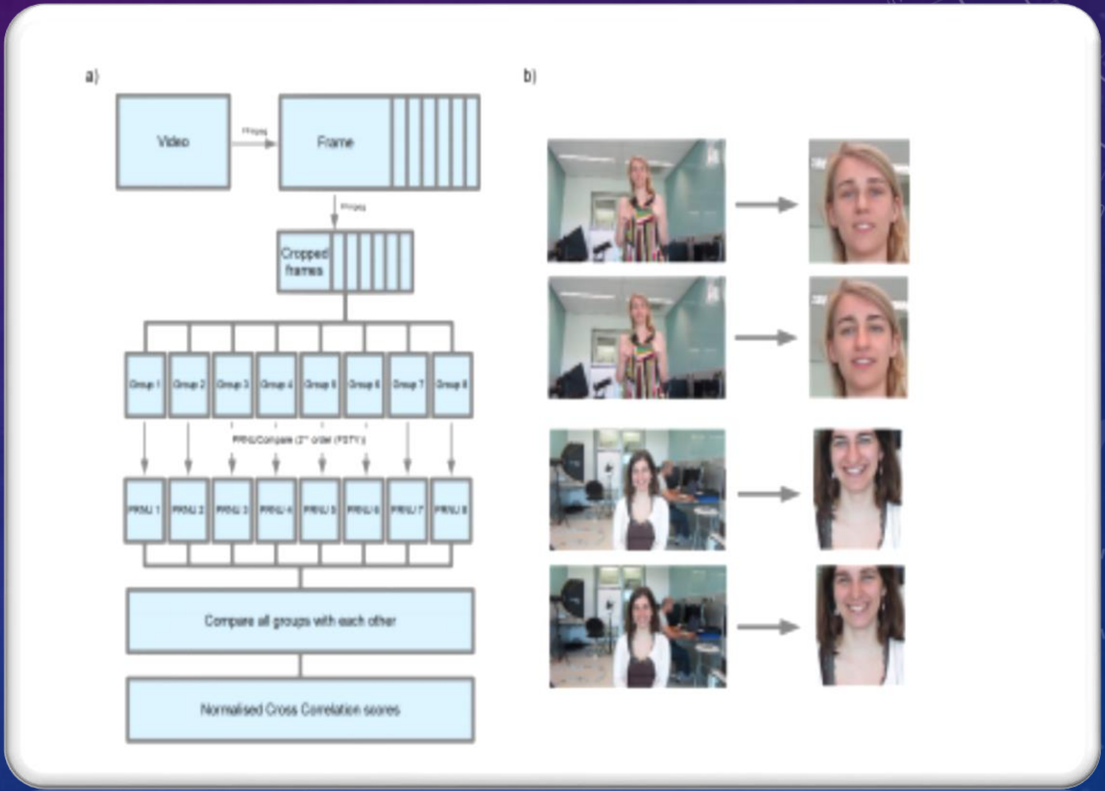


Pre-processing
VGG-19
Capsule Network
Post-processing
Dynamic routing
2D-conv -> stats pooling -> 1D-conv
2D-conv -> stats pooling -> 1D-conv
2D-conv -> stats pooling -> 1D-conv
Real Image Capsule
Fake Image Capsule

# TECHNIQUES TO DETECT DEEPFAKES

- Detection using visual Artifacts(PRNU Patterns)

  - A factory anomaly of light-sensitive sensors in digital cameras causes a unique PRNU noise patterns; acting as a fingerprint for images, this is a popular way to differentiate the digital images.
  - As the swapped face is supposed to modify the original PRNU pattern in the face area of video frames
  - we extract the video frame by frame and then distributed sequentially and evenly in eight groups and calculate an average frame photo response non-uniformity (PRNU) pattern.
  - We sequentially divide the frames into eight classes of identical size and an average PRNU pattern is generated for each class using the second-order (FSTV) method [Baar et al., 2012] with the tool ’PRNUCompare’



# CONCLUSIONS AND FURTHER RESEARCH

## Conclusions

- In this study, we focused on the influence of using different methodologies in the detection of deepfake videos and proposing a homogenous benchmark for follow up work
- In order to create better tools to detect deepfake videos several of the methods above will need to be combined to provide better analysis.
- Combining the bi-spectral analysis of audio along with face warping artifacts within videos are examples of two phases of the mechanism which would begin a step by step process to help eliminate fake videos being posted.
- 2D images and moderate level deepfakes can be identified with image face point technique which accounts for 30% of the fake videos created

## Further Research

- Data drift may be a key factor in identifying anomalies in deepfake videos based on the datasets used to create them. This theory would require further testing and the setup of a GAN network at Pace University
- Incorporating audio bi-spectral analysis into deepfake video detection to distinguish human speech from synthesized speech as an identifier.
- Improving on tools using tensor flow and identifiers such as the random forest algorithm to identify gradient decent within videos as an identifier.