How to Detect Facial Manipulation Image Using CNN

Wang Yang, Jay Xiong,
Liu Yan, Hao Xin, Wei Tao
Baidu X-Lab

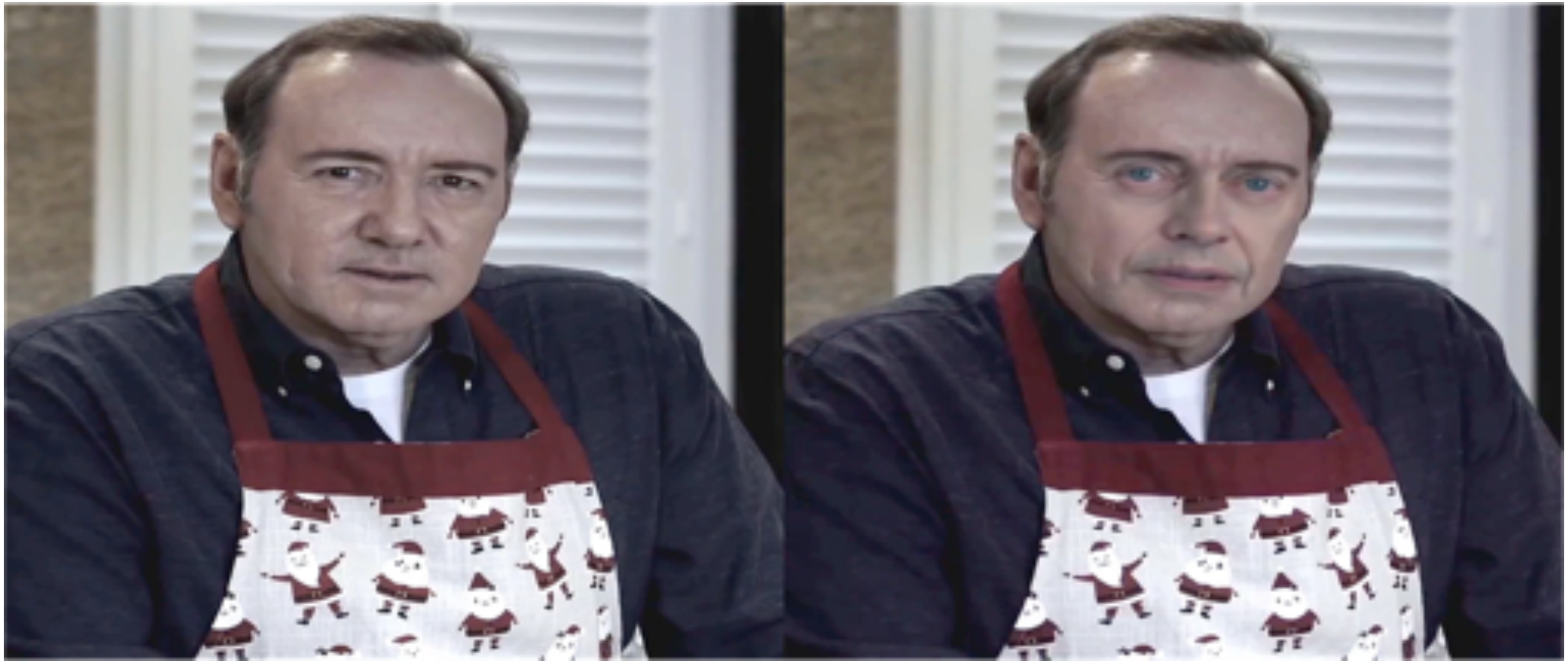
百度安全
有 AI 更 安 全

Baidu Security
Advanced AI , Stronger Security

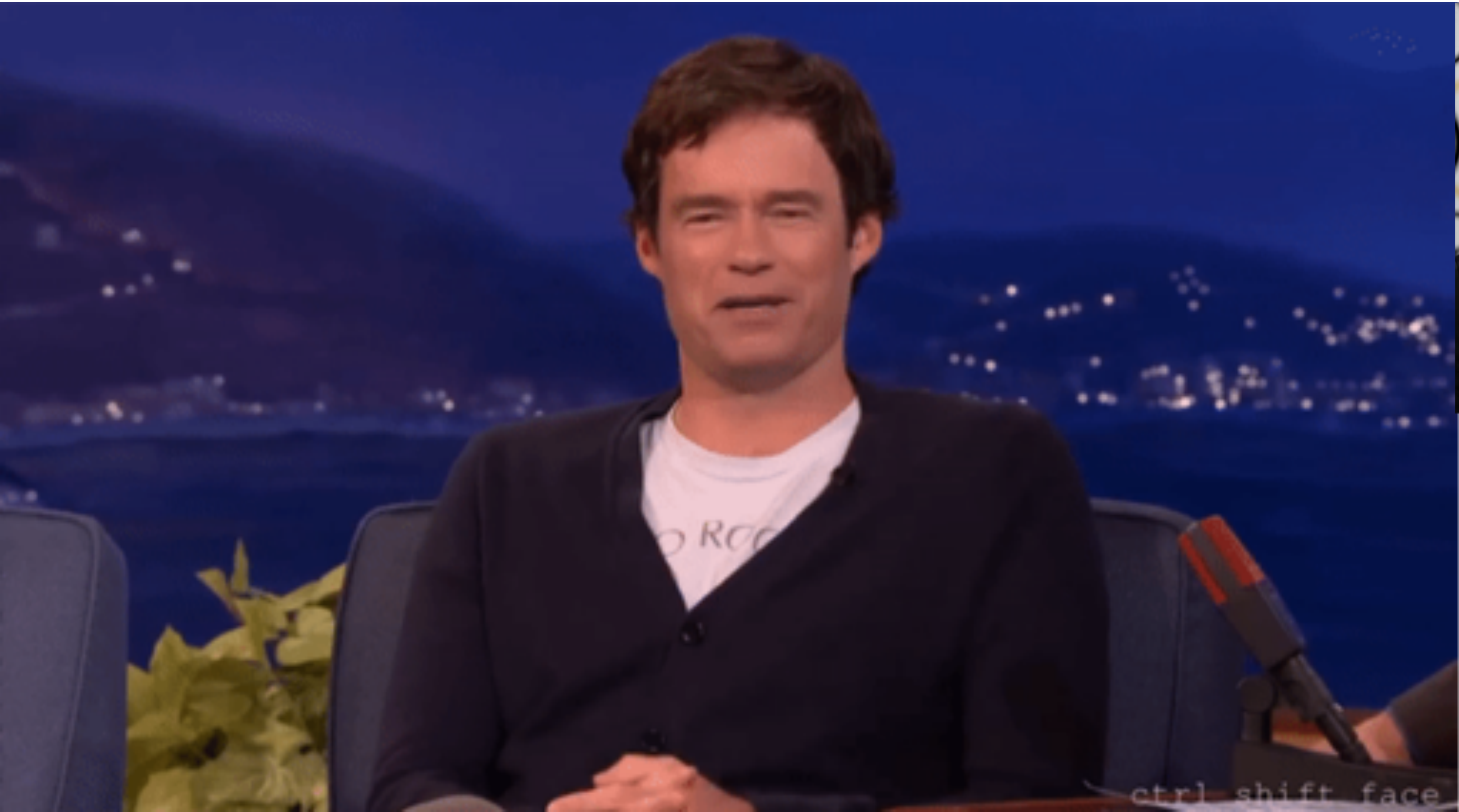
How to Detect Facial Manipulation Image Using CNN

- Facial Manipulation
  - Face swapping
  - Face merging
- A simple and effective CNN
- Face recognition based method

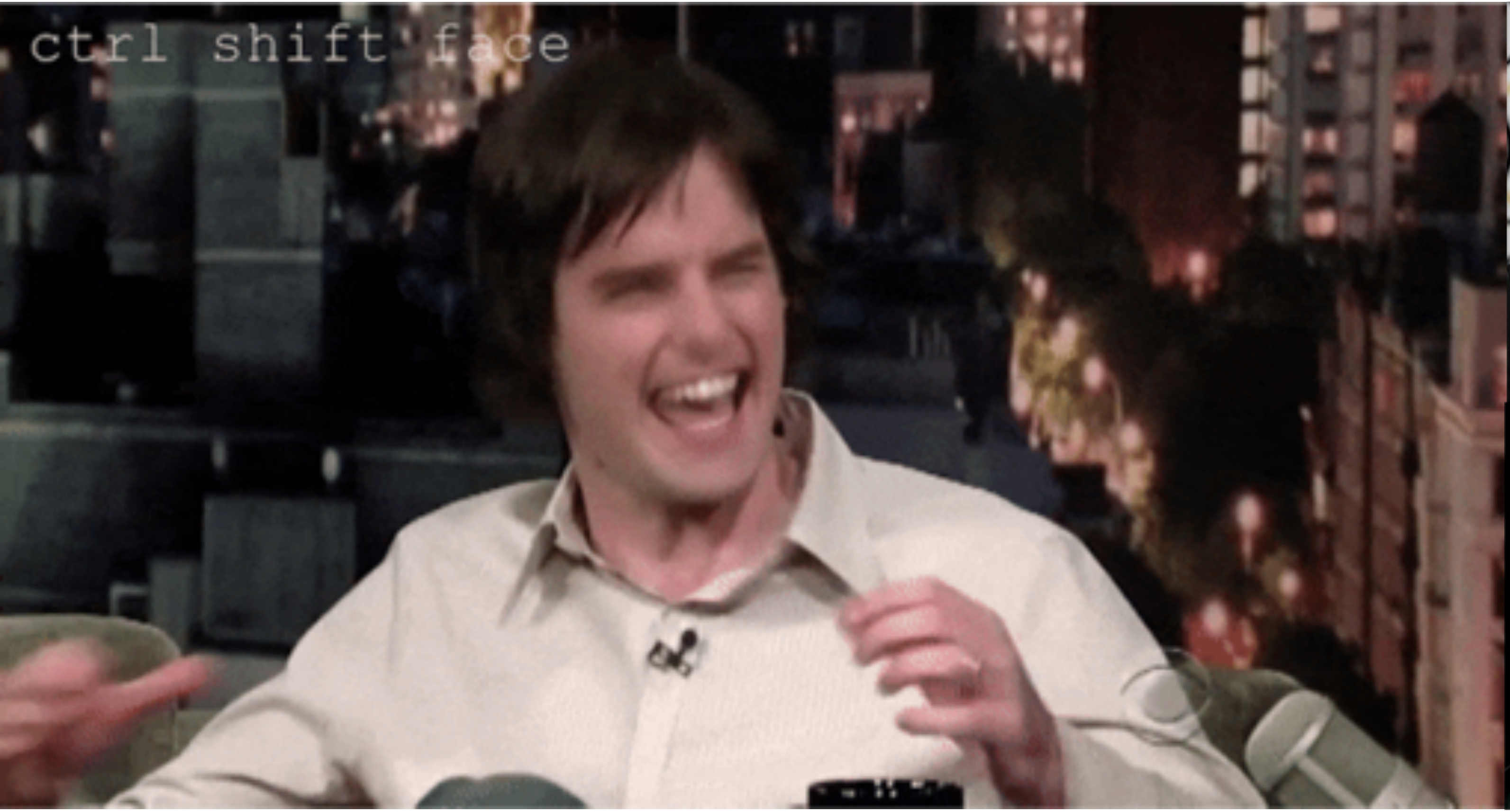
Deepfakes Video



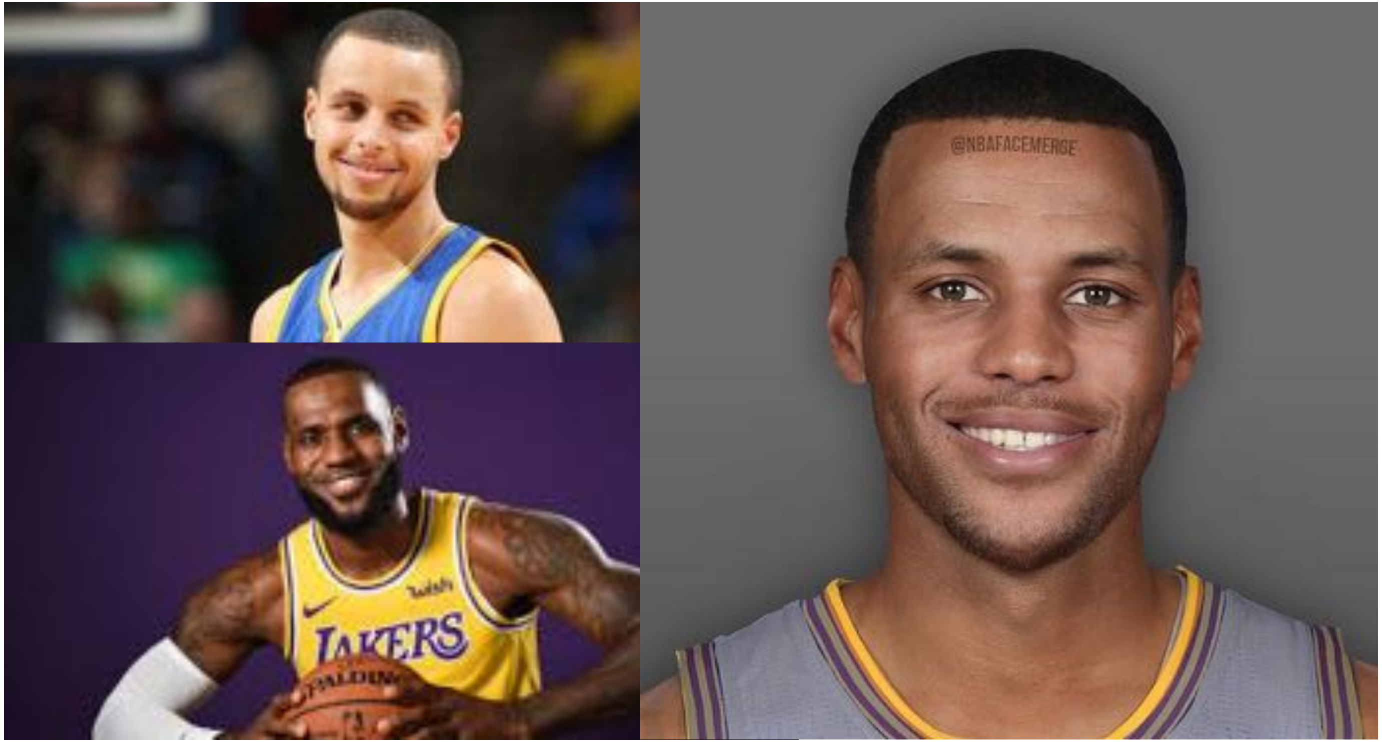
Deepfakes Video



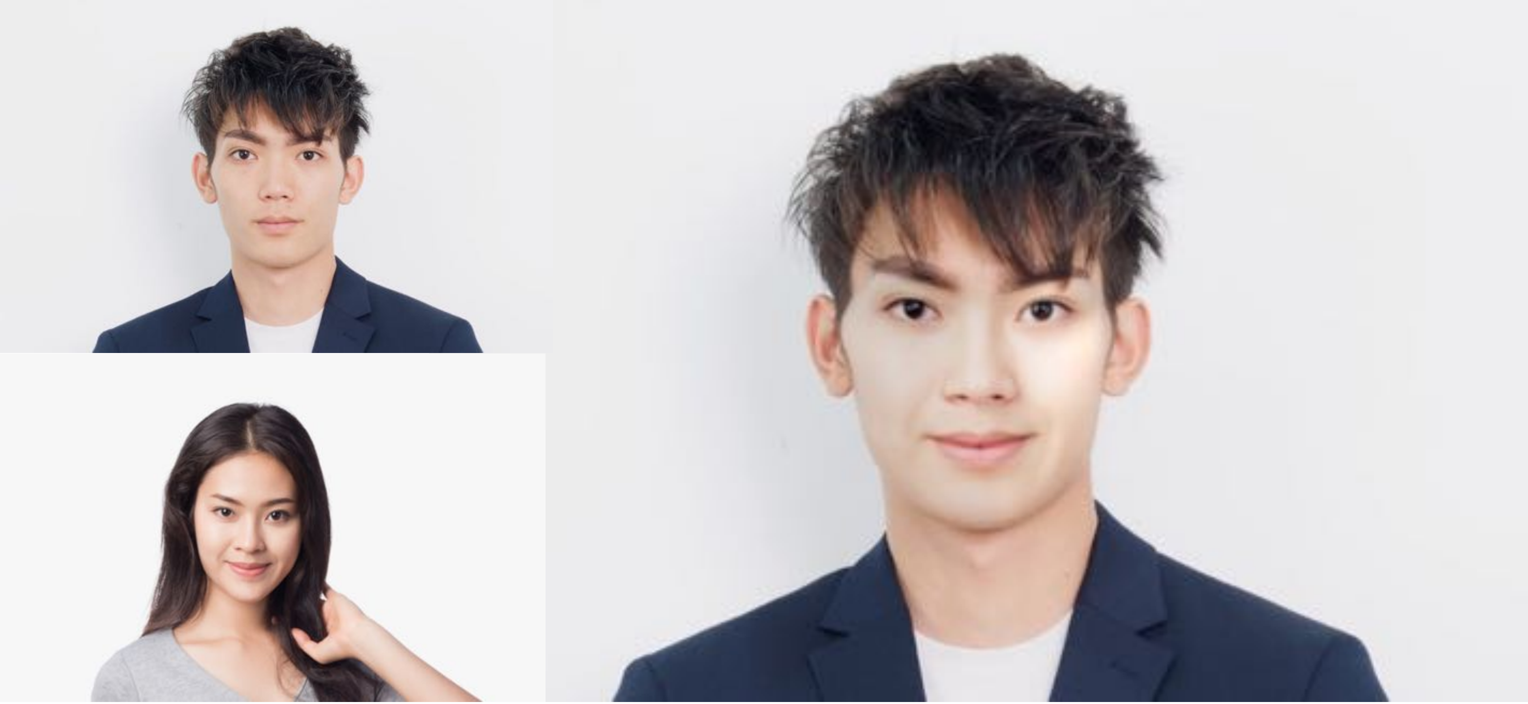
Deepfakes Video



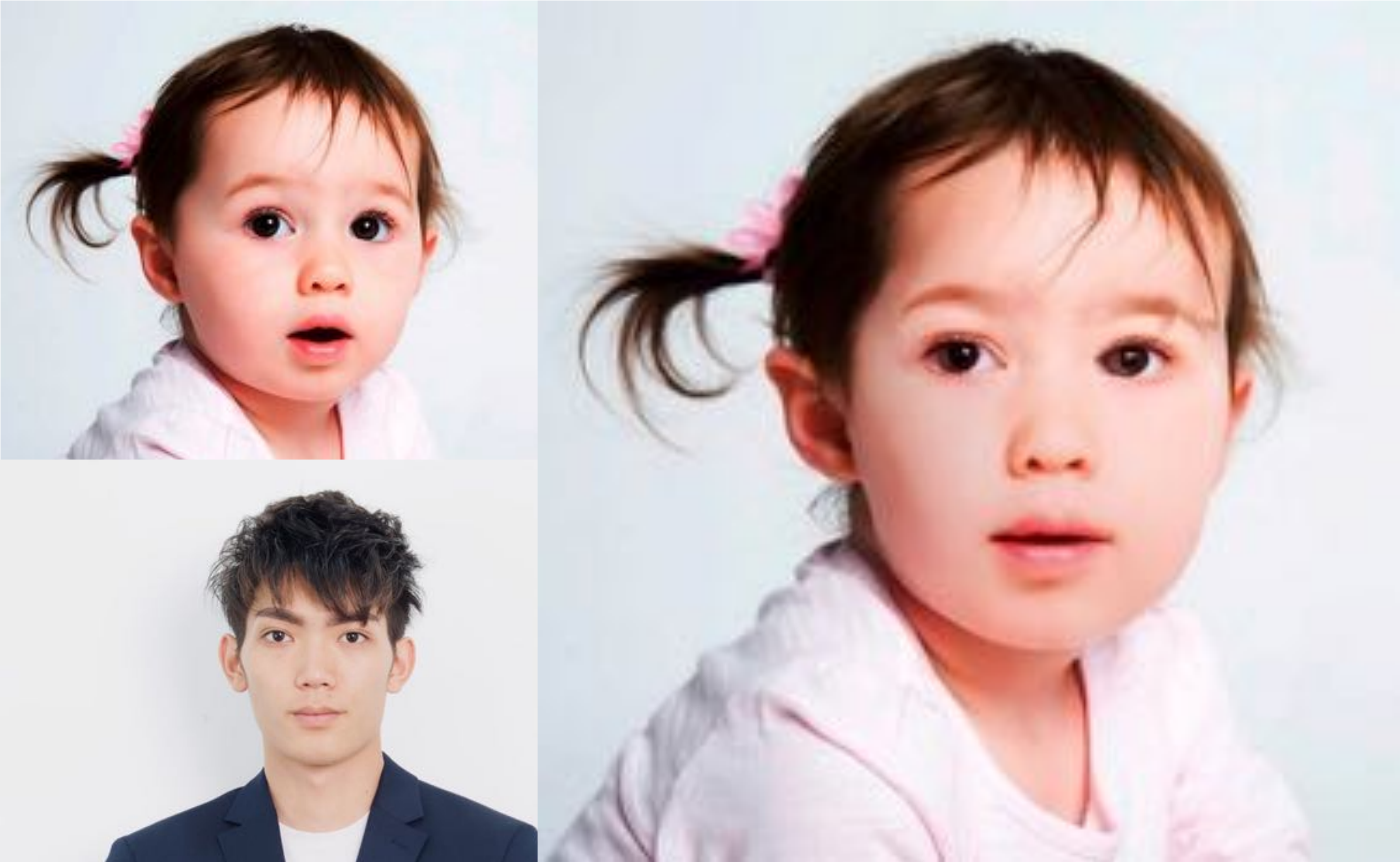
Face Merging Image



Face Merging Image



Face Merging Image

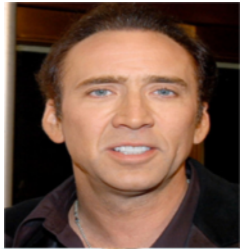


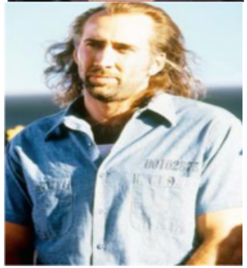
When Face Recognition Systems Meet Deepfakes

Vulnerable face comparison before fake faces

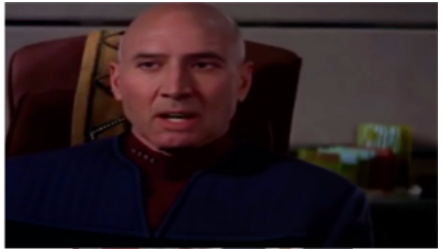
- Microsoft Azure

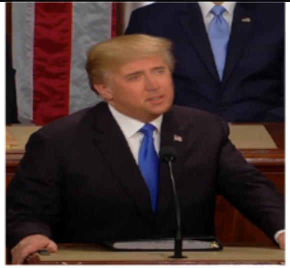
Real





Fake





azure.microsoft.com

Similarity 86.0%

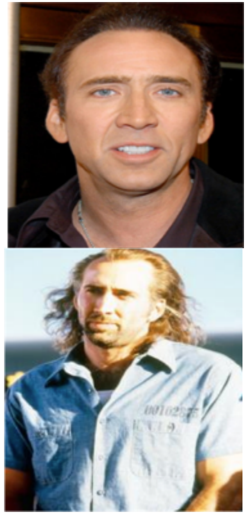
Similarity 70.5%

When Face Recognition Systems Meet Deepfakes

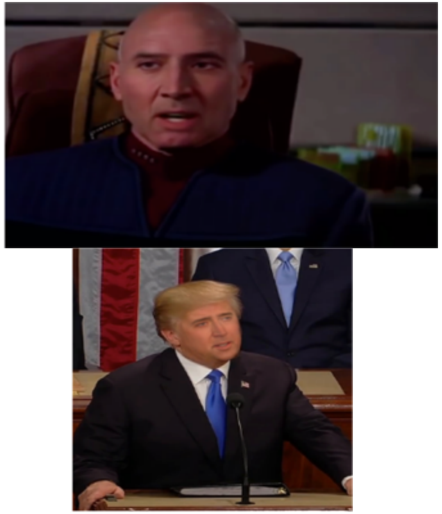
Vulnerable face comparison before fake faces

- Amazon AWS

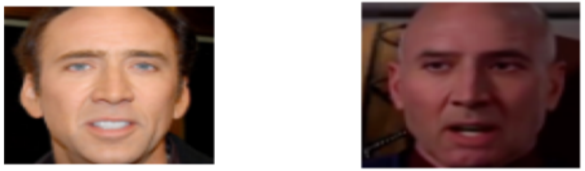
Real



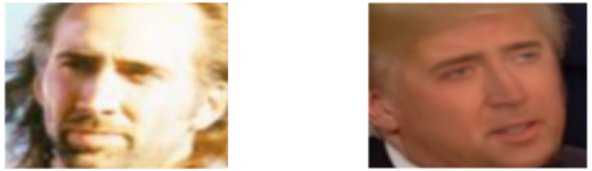
Fake



aws.amazon.com



Similarity 95.1%

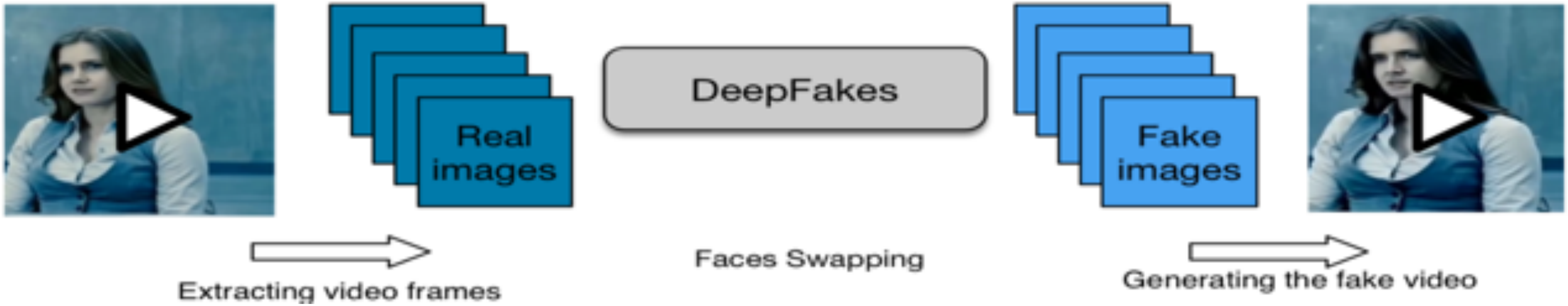


Similarity 87.3%

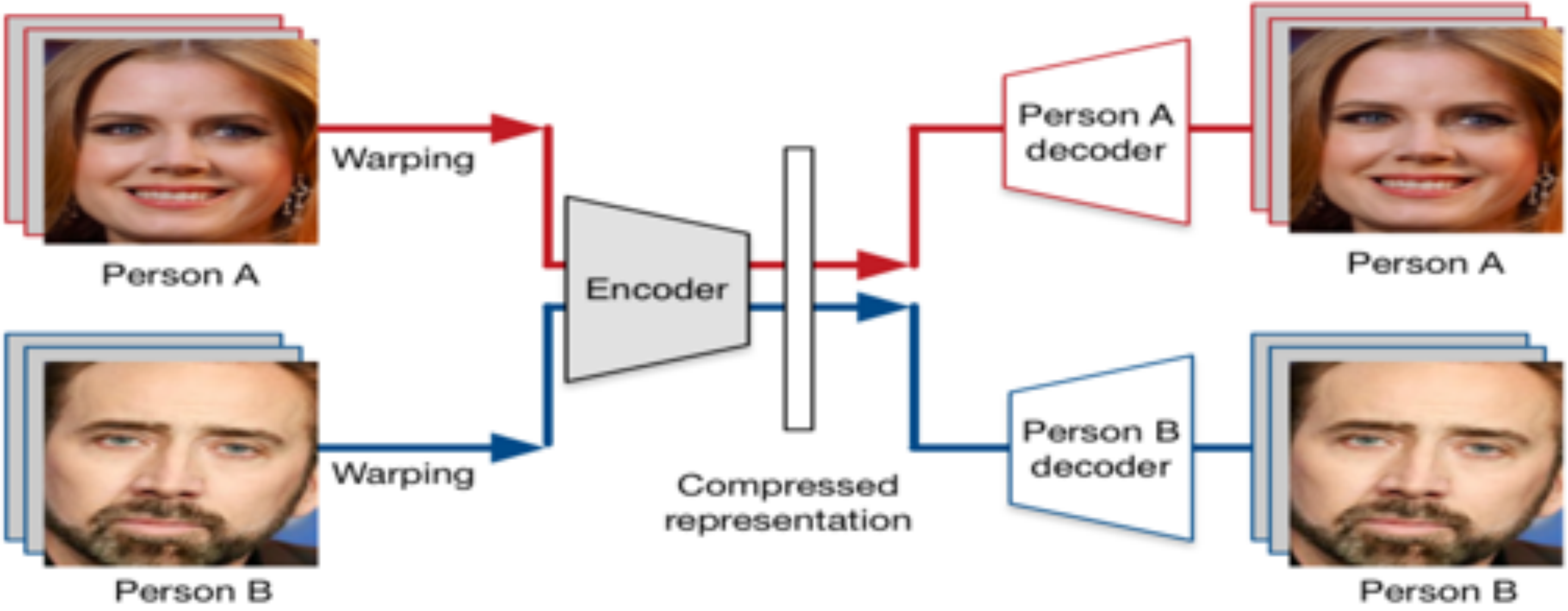
Face Swapping Video Generation

Characteristics

- Swap victim's face in every frame independently
- Not End2End
- Only manipulate central face area
- Autoencoder

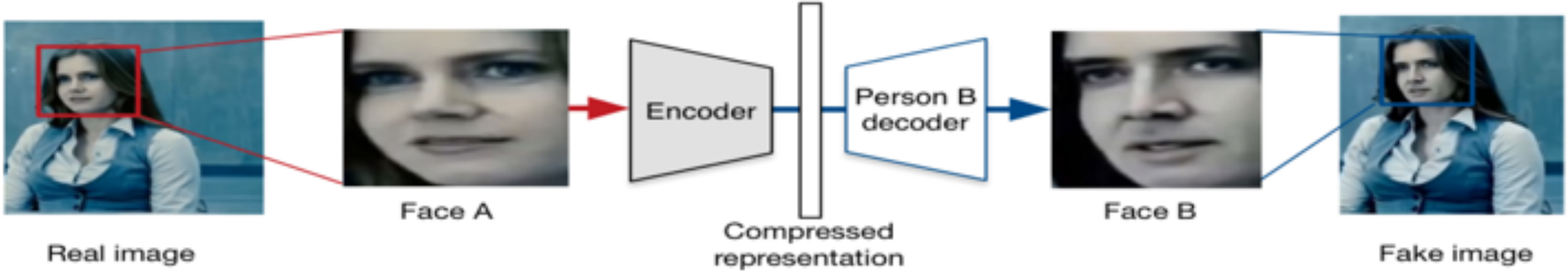


Deepfakes Training Phase



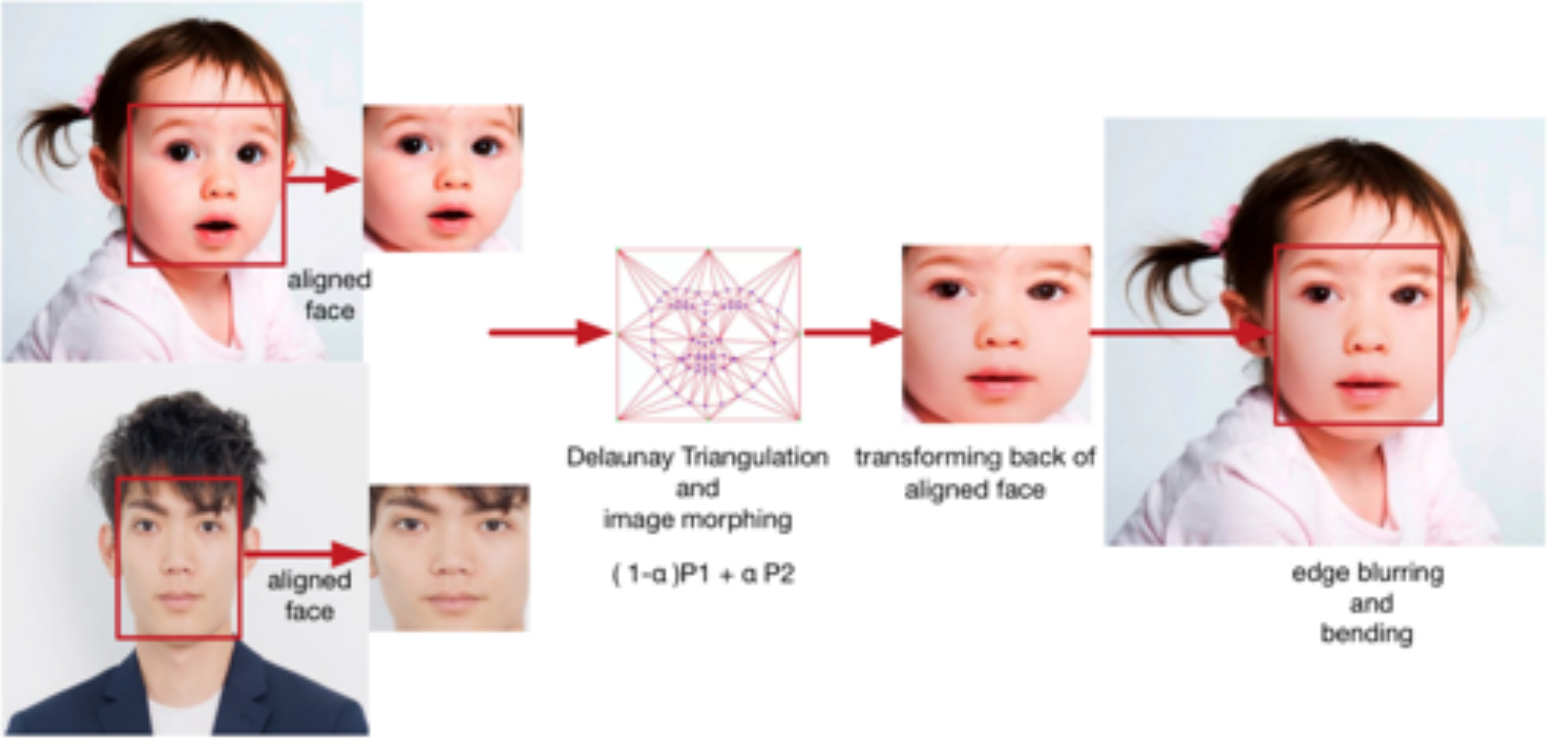
Deepfakes Generation Phase

- Convert
  - Person A Encoder -> Person B Decoder
- Merge back
  - Gaussian Blur/Color Average
  - Poisson Image Editing



Generating Face Merging Image

- No training required


aligned face
aligned face
Delaunay Triangulation and image morphing
(1-α)P1 + α P2
transforming back of aligned face
edge blurring and bending

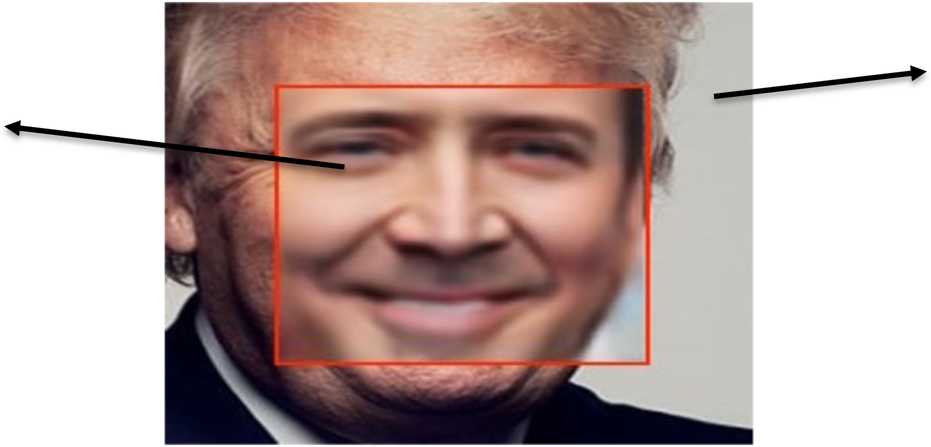
How to Detect Facial Manipulation Image Using CNN

- Facial Manipulation
- **A simple and effective CNN**
  - capturing low-level features of the images
- Face recognition based method

A Simple and Effective CNN

Design purpose

- Input contains marginal(background) information.
- Capture low-level features of the images.



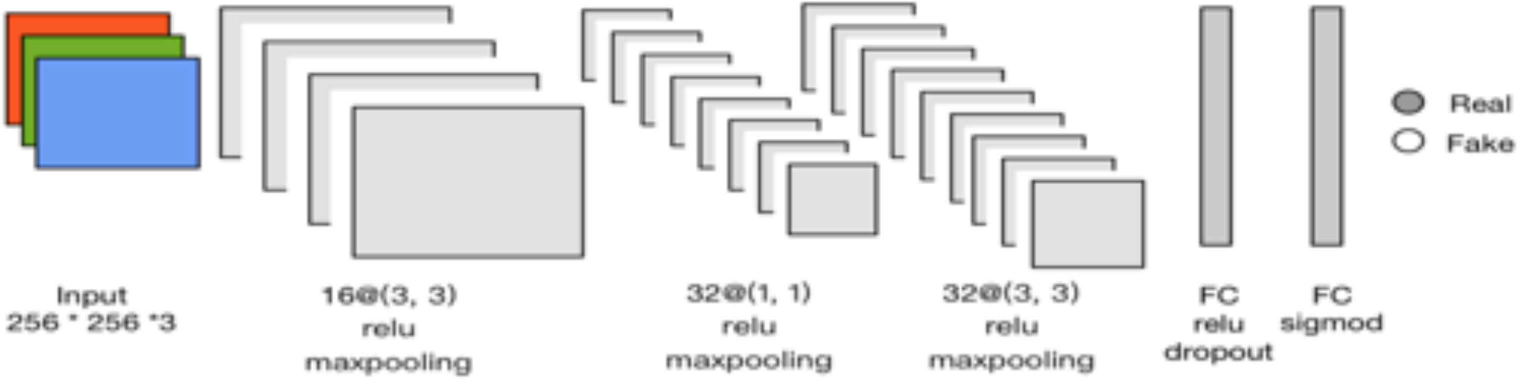
face area

margin
(background)

A Simple and Effective CNN

Characteristics

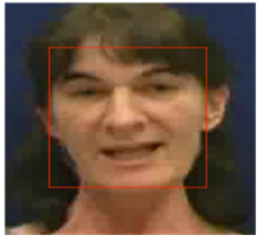
- 3 convolution layers
- Accuracy rate
  - accuracy 98% for Deepfakes
  - accuracy 92% for face merging



A Simple and Effective CNN

Training

- Dataset of Deepfakes
  - VidTIMIT and Youtube
  - 142,458 fake faces and 141,932real faces
  - low quality and high quality images
- Dataset of face merge
  - LFW and VGGFACE
  - 11,340 fake faces and 11,839 real faces
  - Baidu face merging API
- Cropped faces
  - with face landmark detector MTCNN
  - obtain 1.5 scaled bounding box
- Augmented data
  - horizontal flipping
  - randomly zooming
  - shearing transformations



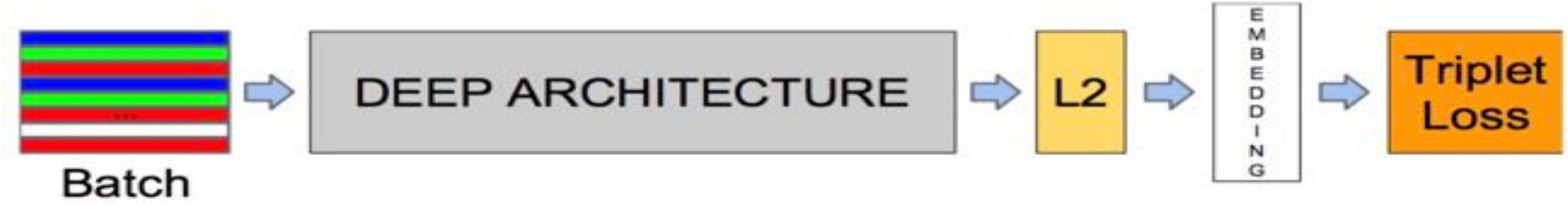
How to Detect Facial Manipulation Image Using CNN

- Facial Manipulation
- A simple and effective CNN
- Face recognition based method
  - capturing high-level features of faces

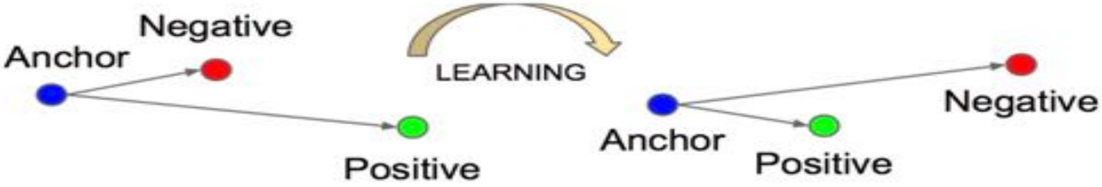
What is FaceNet?

Characteristics

- SOTA CNN for face recognition
- Model structure



- Triple Loss

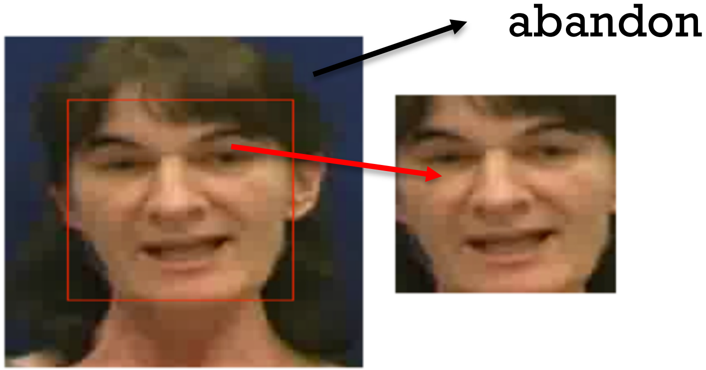


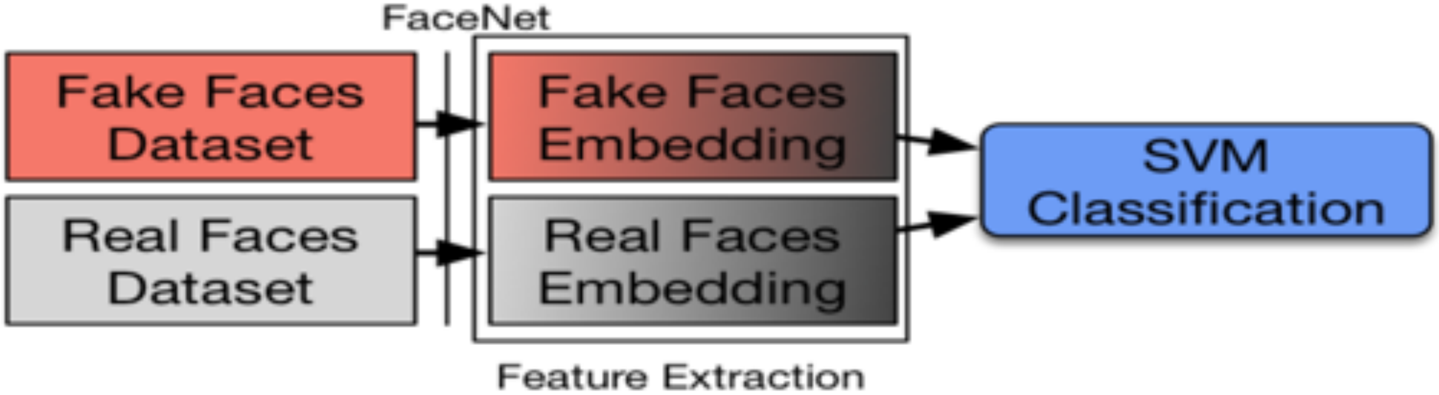
FaceNet: A unified embedding for face recognition and clustering

A FaceNet based SVM Classifier

Training

- Central face area
  - No margin/background
  - Only face area

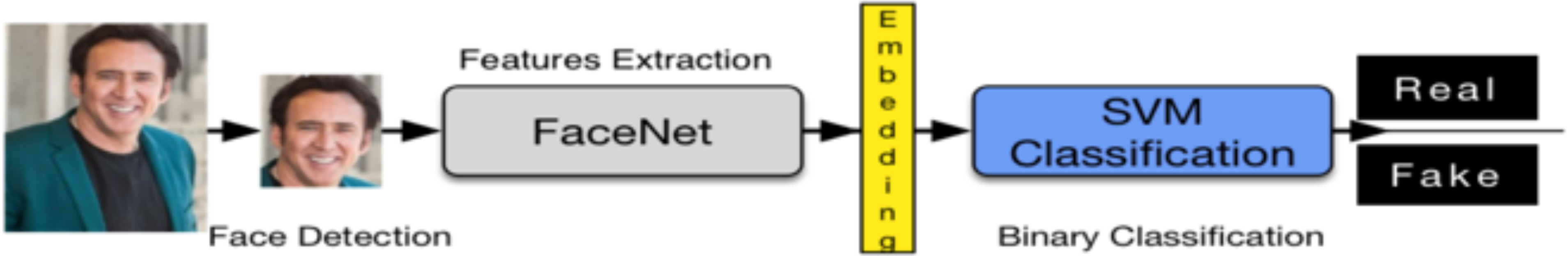




A FaceNet based SVM Classifier

Characteristics

- FaceNet used for extracting face features
- SVM for binary classification
- Accuracy rate
  - accuracy 93% for Deepfakes
  - accuracy 64% for face merging



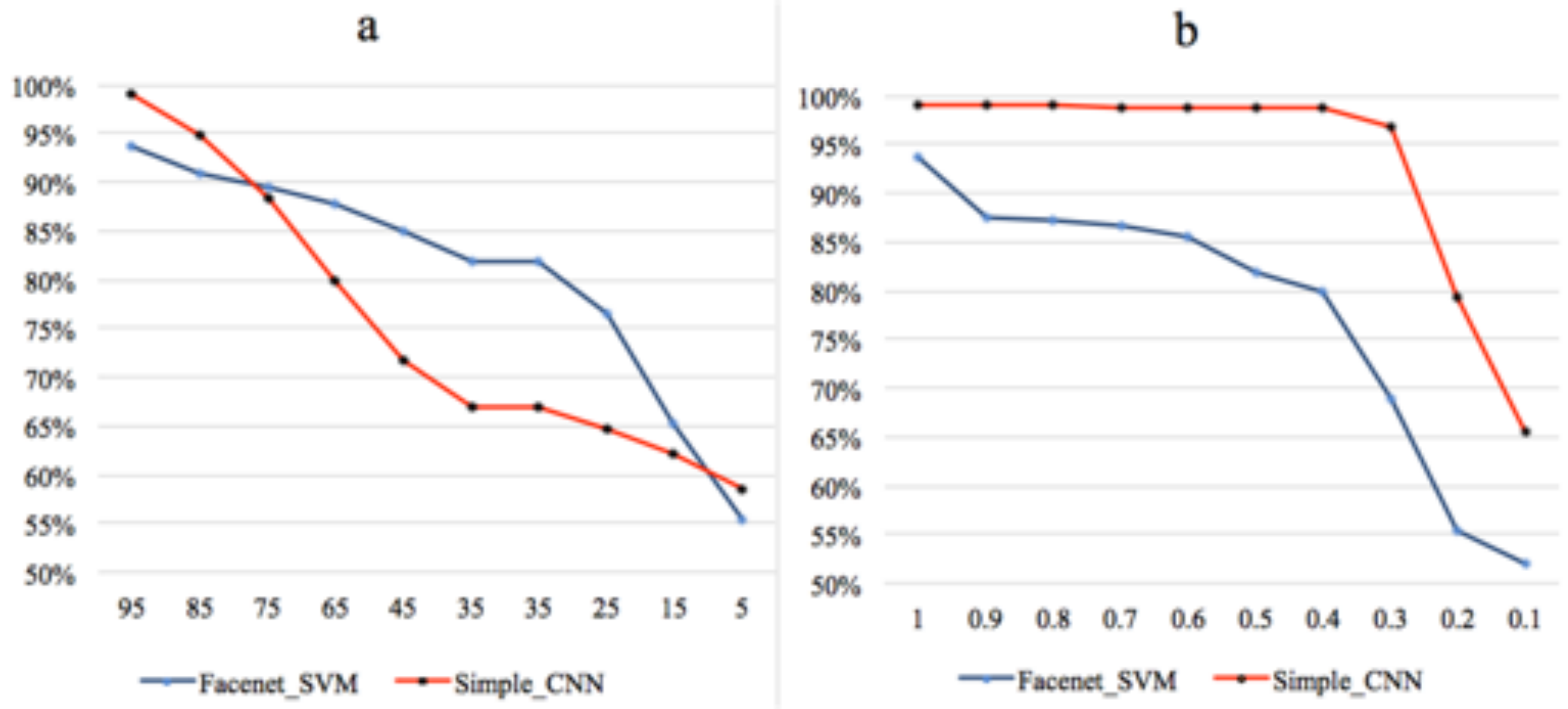
A Simple and Effective CNN

Accuracy rate on Deepfakes: 98%

A FaceNet based SVM Classifier

Accuracy rate on Deepfakes: 92%

Performance Under Compression and Resizing



Summary

- CNN for image classification
  - A simple architecture can work well.
  - catching low-level features: contours, edges…
  - other models for detecting Deepfakes

Meso-4	MesoInception-4	VGG16	Inception
94%	98%	96%	86%

- A FaceNet based SVM classifier
  - using FR to catch features of fake faces
  - using SVM for binary classification
  - 64% accuracy rate for the misclassification set from the simple CNN based classifier

Thank You!
Q&A