# Spoofing Cybersecurity

## Arming your organization against deepfakes and AI-synthesized data

**Adel Abusitta**


GOLDEN GLOBE AWARDS

# Agenda

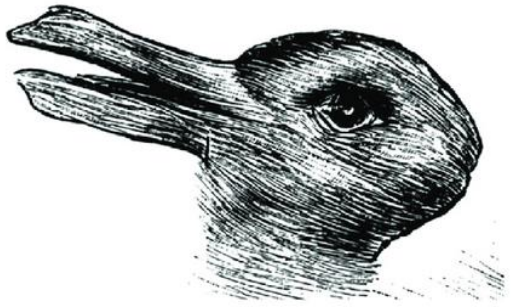
- The State of Deepfakes
- Generative Adversarial Networks and Deepfakes
- How to detect Deepfackes
- How to prepare an organization against Deepfakes

# Why does Deepfakes matter?

## Deepfakes matter because:

**Believability**: If we see and hear something with our own eyes and ears, we believe it to exist or to be true, even if it is unlikely.

- The brain's visual system can be targeted for misperception, in the same way optical illusions and bistable figures trick our brains.



Jastow rabbit-duck

Rubin vase-faces

Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat?. *Business Horizons*, *63*(2), 135-146.

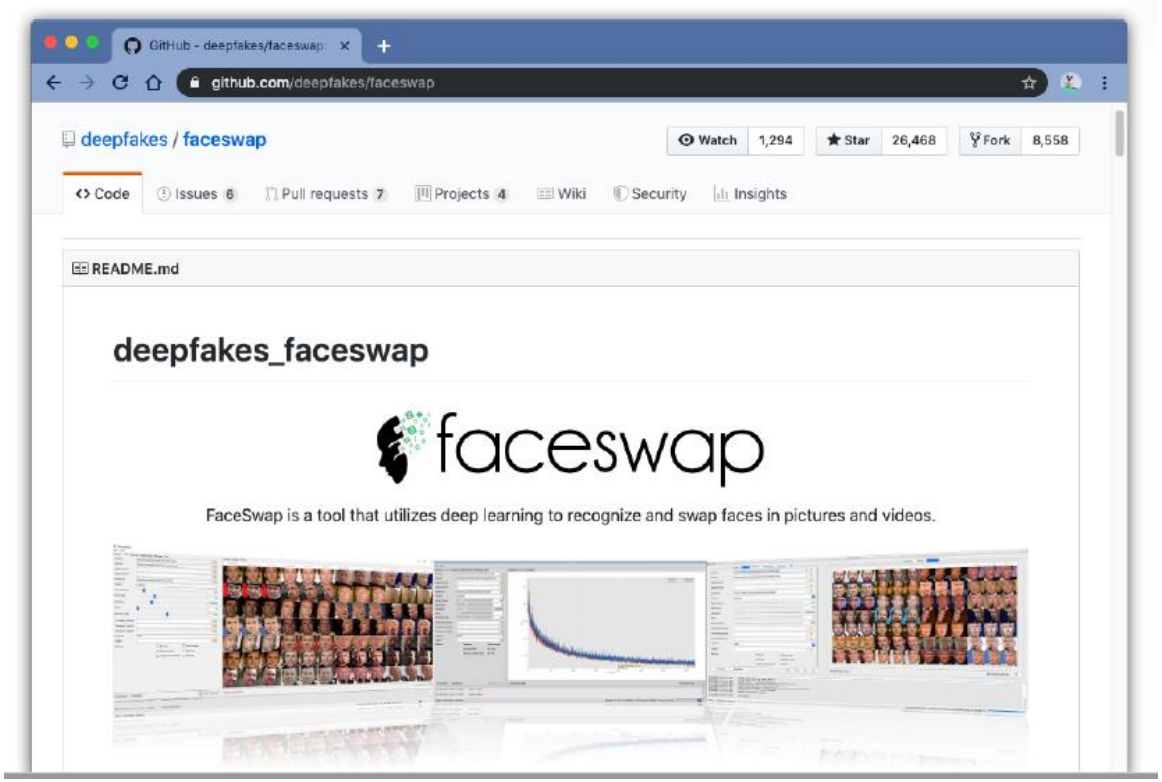
# Why does Deepfakes matter?

## Deepfakes matter because:

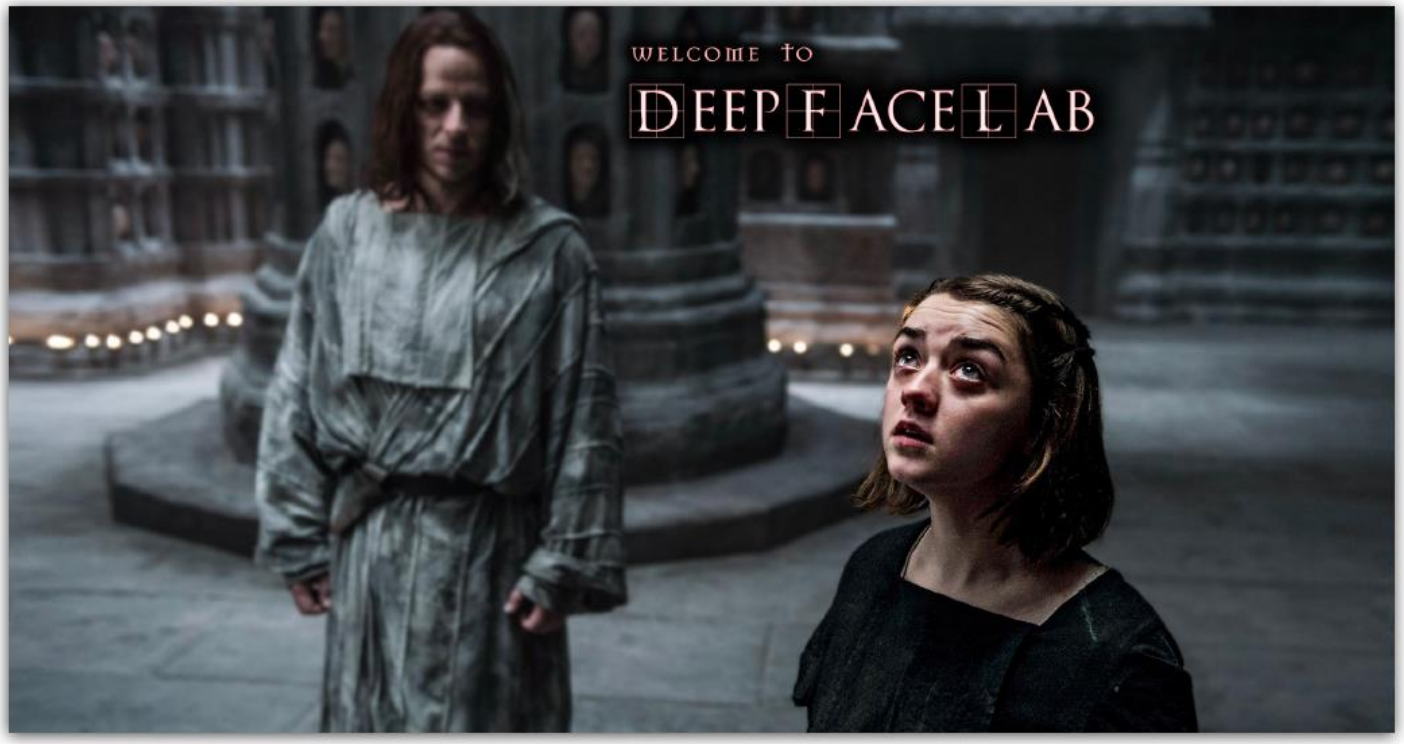
**Accessibility:** the technology of today and tomorrow, will allow all of us to create fakes that appear real, without a significant investment in training, data collection, hardware and software.

# Faceswap

Open Source multi-platform Deepfakes software.



# DeepFaceLab



# Zao ,

The popular Chinese app for mobile devices lets users place their faces into scenes from movies and TV shows, for free.

# How do deepfakes work?

## How do deepfakes work?

Consider below deepfake featuring Jim Carrey and Alison Brie:

**Original** showing Alison Brie

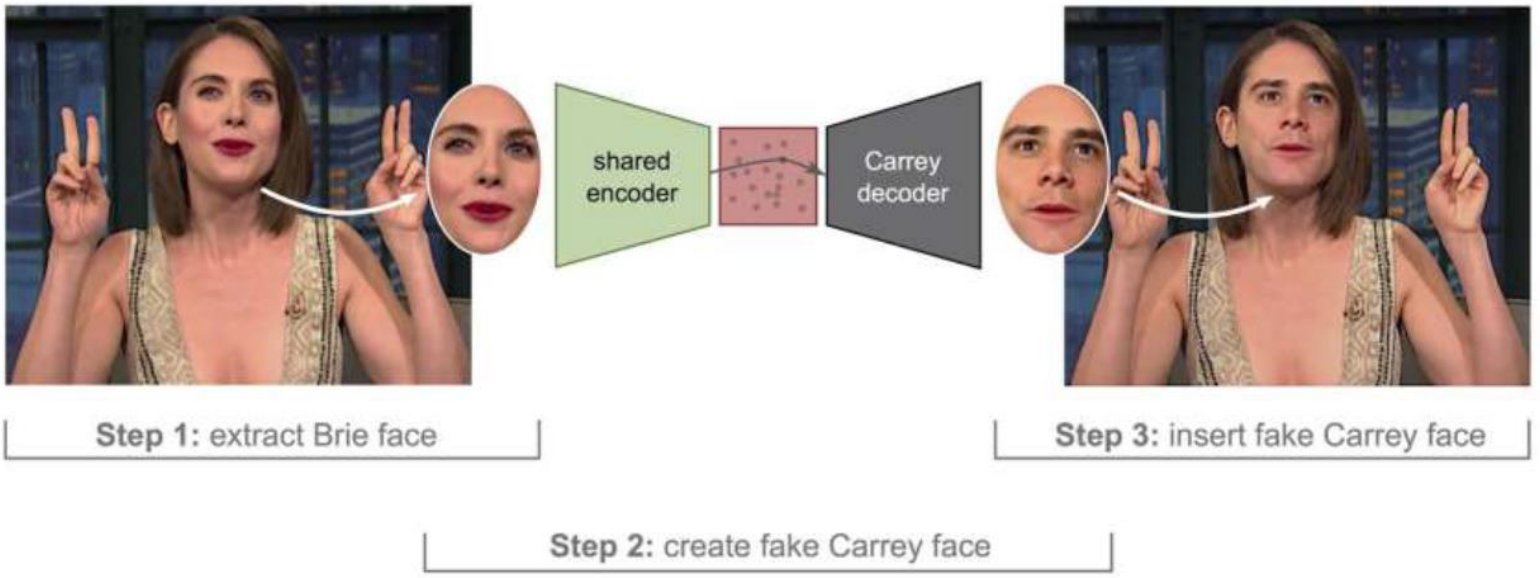
**Deepfake** showing Jim Carrey instead of Brie

The original Alison Brie video: https://www.youtube.com/watch?v=QBmYDzLhWoY
The deepfake with Jim Carrey: https://www.youtube.com/watch?v=b5AWhh6MYCg

Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat?. *Business Horizons*, *63*(2), 135-146.

# How do deepfakes work?

Many deepfakes are created by a three-step procedure:




Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat?. *Business Horizons*, *63*(2), 135-146.

# How do deepfakes work?

- **Step 1:** The image region showing Brie's face is extracted from an original frame of the video. This image is then used as input to a deep neural network (DNN)

- **Step 2:** The DNN automatically generates a matching image showing Carrey instead Brie

- **Step 3:** This generated face is inserted into the original reference image to create the deepfake

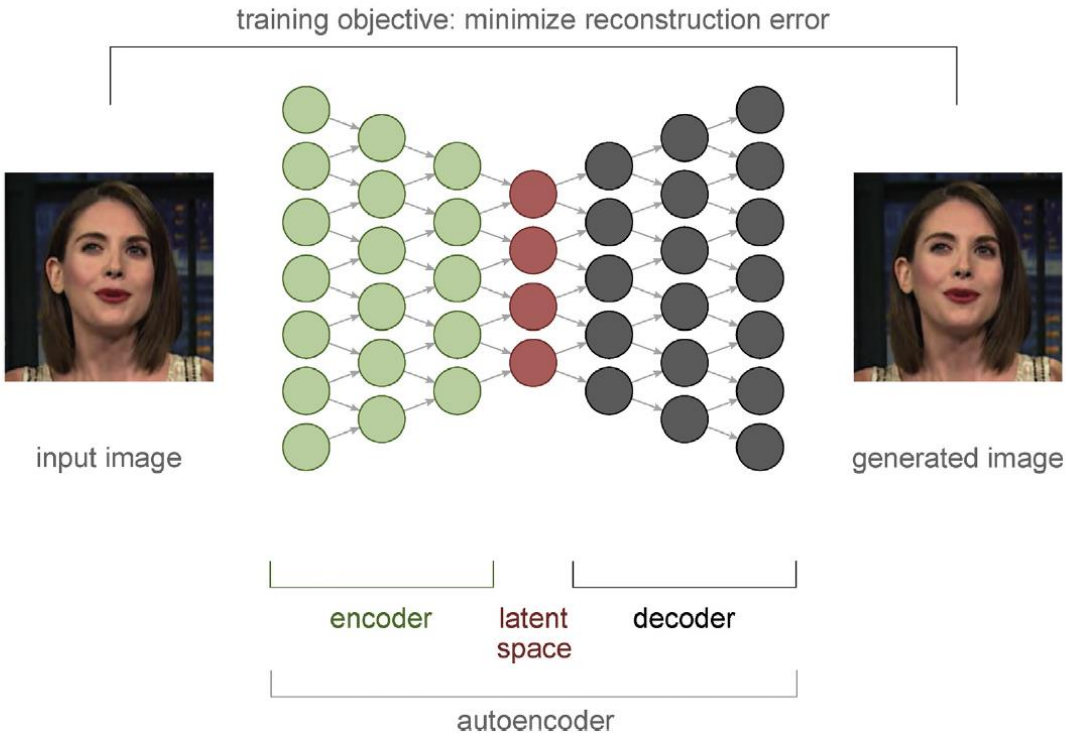
# How do deepfakes work?

- Deepfakes are commonly created using a specific deep network architecture known as **autoencoder**.

- Autoencoders consist of three subparts:
  - ***an encoder*** (recognizing key features of an input face)
  - ***a latent space*** (representing the face as a compressed version)
  - ***a decoder*** (reconstructing the input image with all detail)

# How do deepfakes work?

- **Autoencoder:** a DNN architecture commonly used for generating deepfakes.

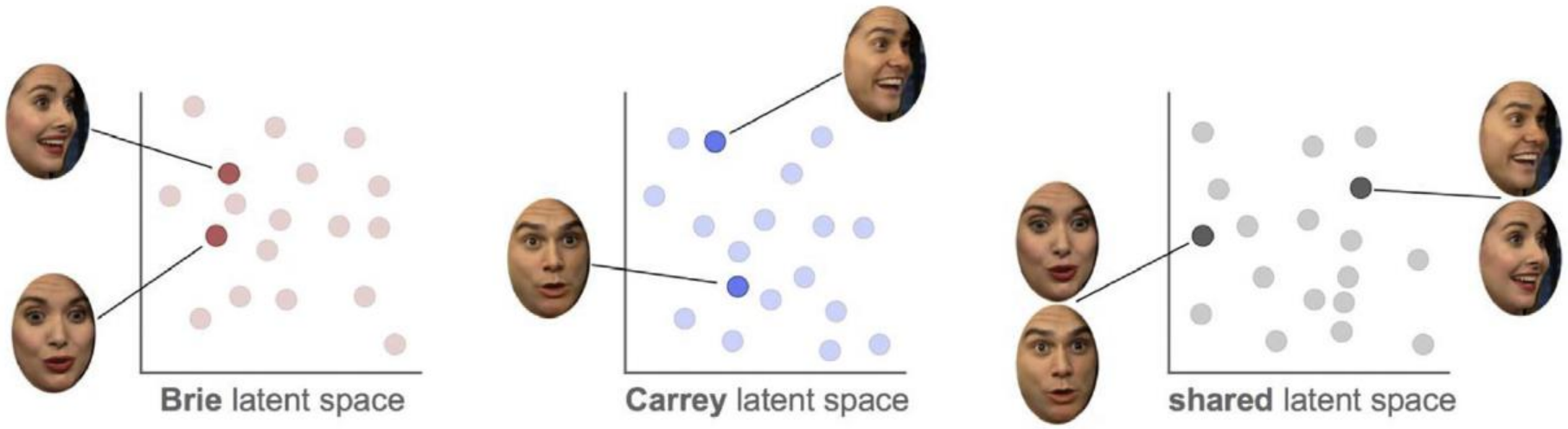



Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat?. *Business Horizons*, *63*(2), 135-146.

# How do deepfakes work?

## The deepfake trick

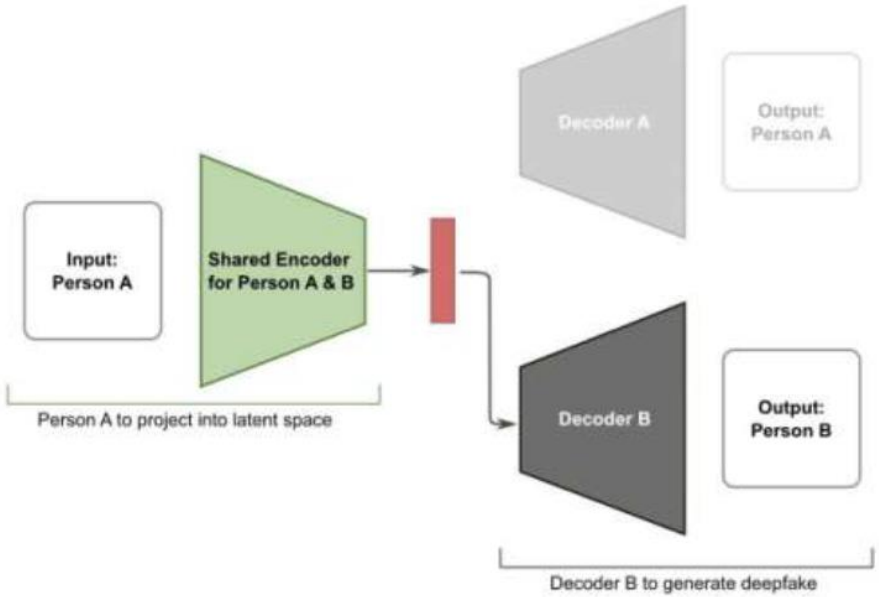
Using the same encoder and hence latent space representation for images of two separate people




Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat?. *Business Horizons*, *63*(2), 135-146.

# How do deepfakes work?

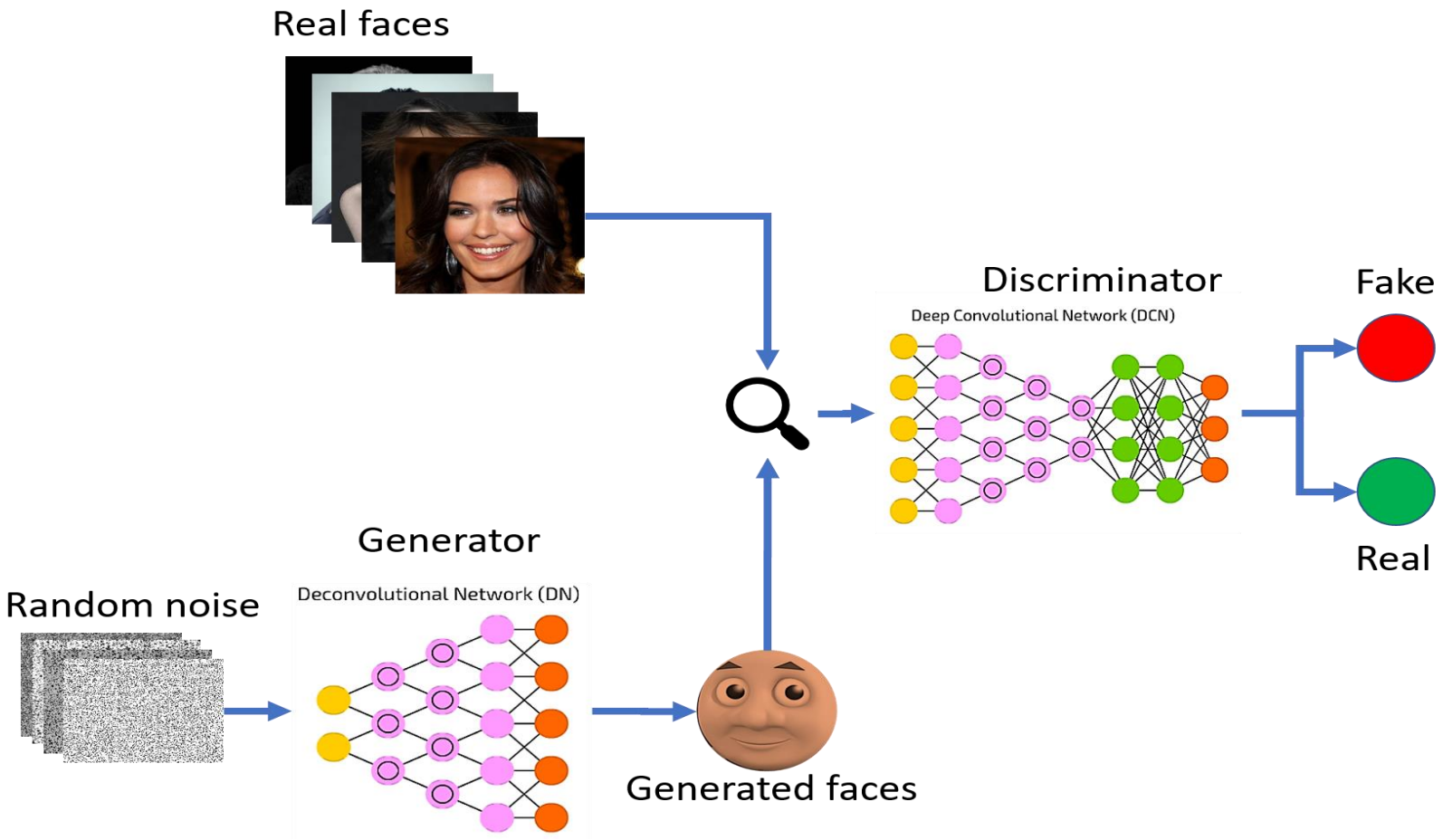
- A **shared encoder** is key to creating novel facial images of a target person that exhibit the same emotional expression, head posture, etc. as the original facial characteristics.
- This new image can then be doctored back into the original image to create a fake scene.




Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat?. *Business Horizons*, *63*(2), 135-146.

# Generative Adversarial Networks and Deepfakes

# Generative Adversarial Networks [Goodfellow et al., 2014]




Source: https://www.freecodecamp.org/news/an-intuitive-introduction-to-generative-adversarial-networks-gans-7a2264a81394/

# Generative Adversarial Networks [Goodfellow et al., 2014]

**Algorithm 1** Minibatch stochastic gradient descent training of generative adversarial nets. The number of steps to apply to the discriminator, $k$, is a hyperparameter. We used $k = 1$, the least expensive option, in our experiments.

**for** number of training iterations **do**

  **for** $k$ steps **do**

  - Sample minibatch of $m$ noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
  - Sample minibatch of $m$ examples $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ from data generating distribution $p_{\text{data}}(\boldsymbol{x})$.
  - Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[ \log D\left(\boldsymbol{x}^{(i)}\right) + \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right) \right].$$

  **end for**

  - Sample minibatch of $m$ noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
  - Update the generator by descending its stochastic gradient:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right).$$

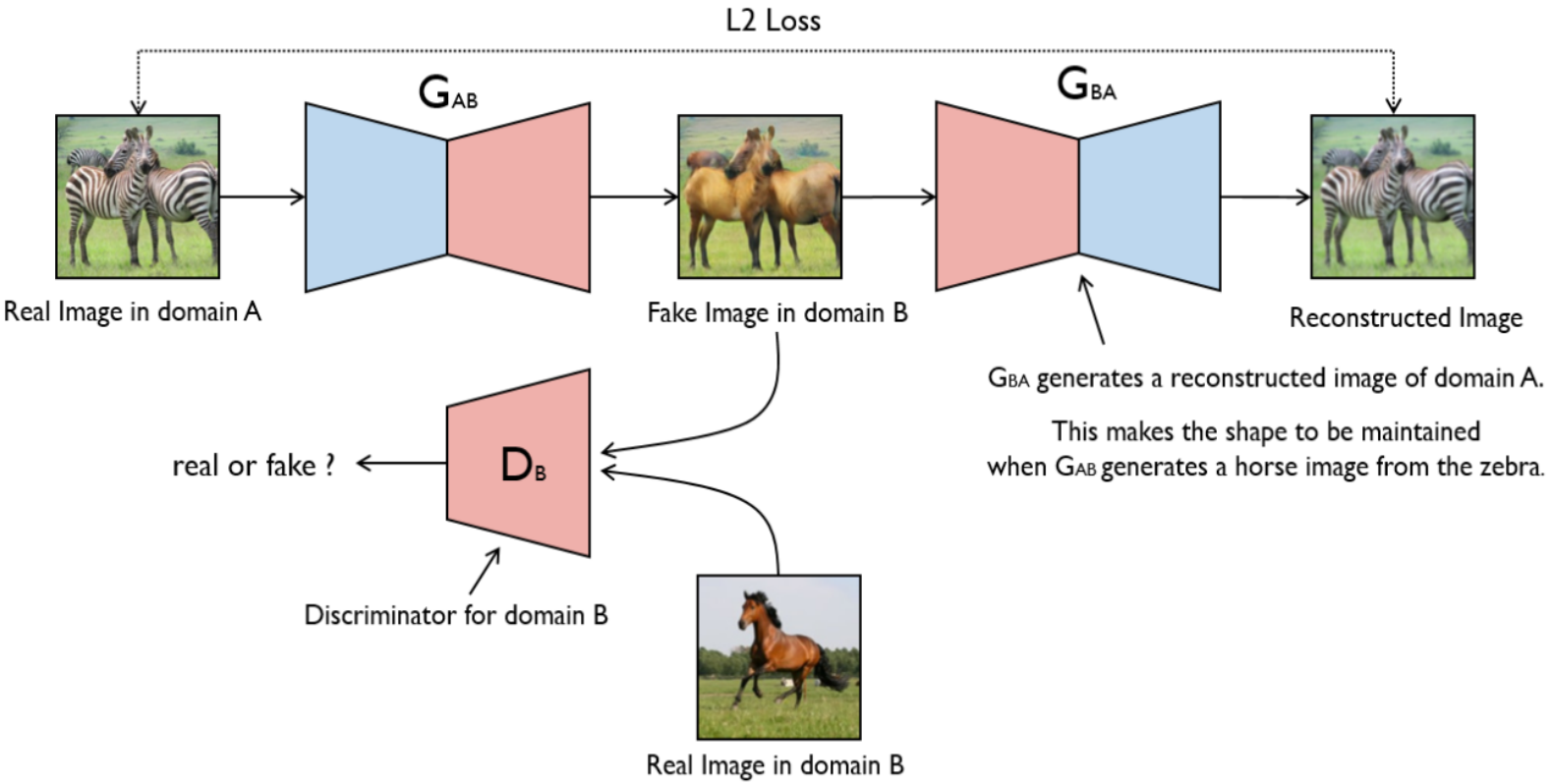
**end for**

The gradient-based updates can use any standard gradient-based learning rule. We used momentum in our experiments.



Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial networks. *arXiv preprint arXiv:1406.2661*.

# CycleGAN




Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223-2232).

https://towardsdatascience.com/image-to-image-translation-using-cyclegan-model-d58cfff04755

# GANs for good

## Comprehensive generative objectives

Learning to Anonymize Faces for Privacy Preserving Action Detection, Ren et al. 2018



Source: Prof. Esma Aïmeur's slide (fake news)

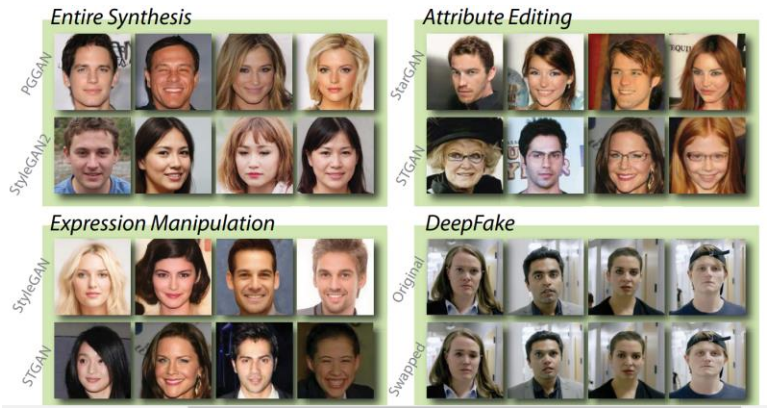
# GAN-based Attacks: four types

i) Entire face synthesis

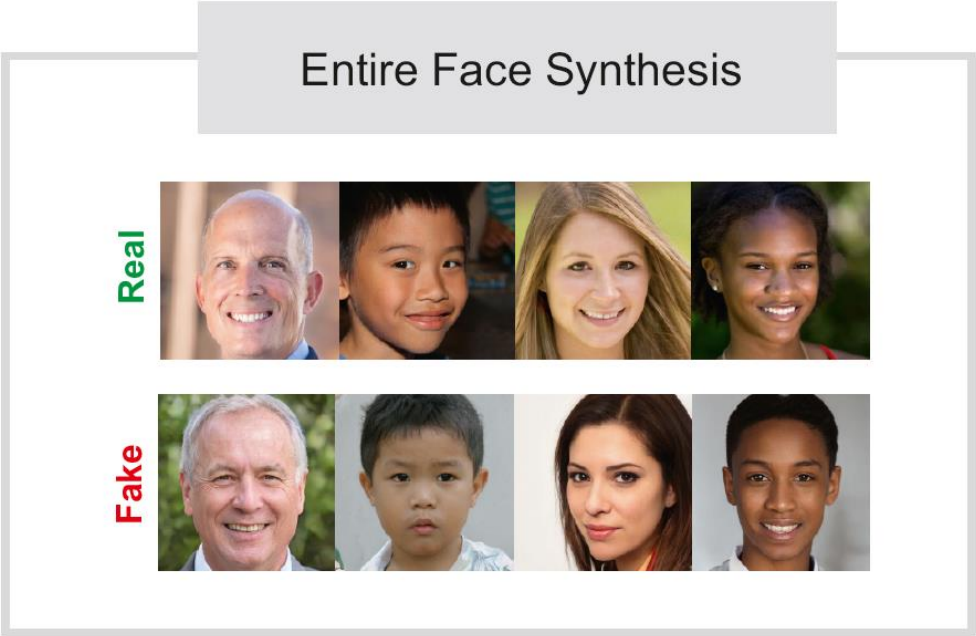
ii) Identity swap (DeepFakes)

iii) Attribute manipulation

iv) Expression swap.

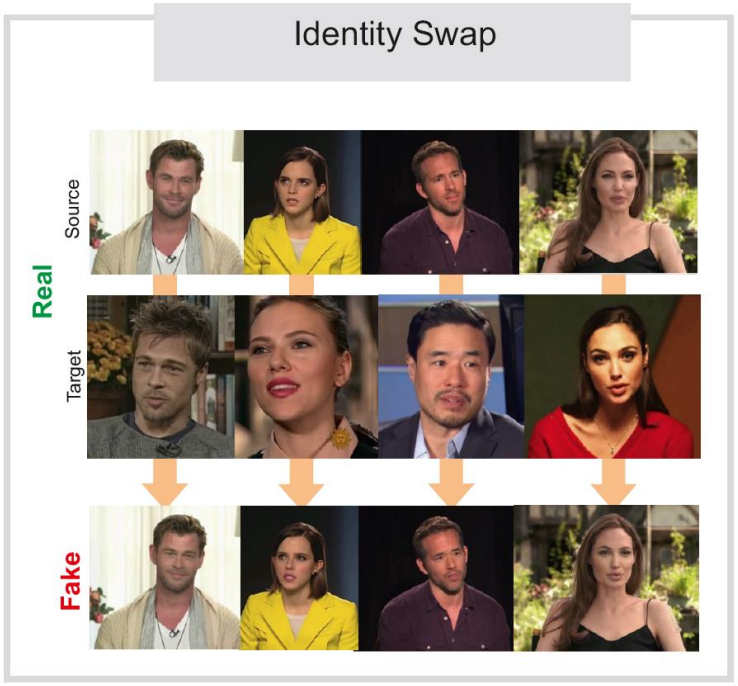


# Entire Face Synthesis




Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, *64*, 131-148.

# Identity swap (DeepFakes)



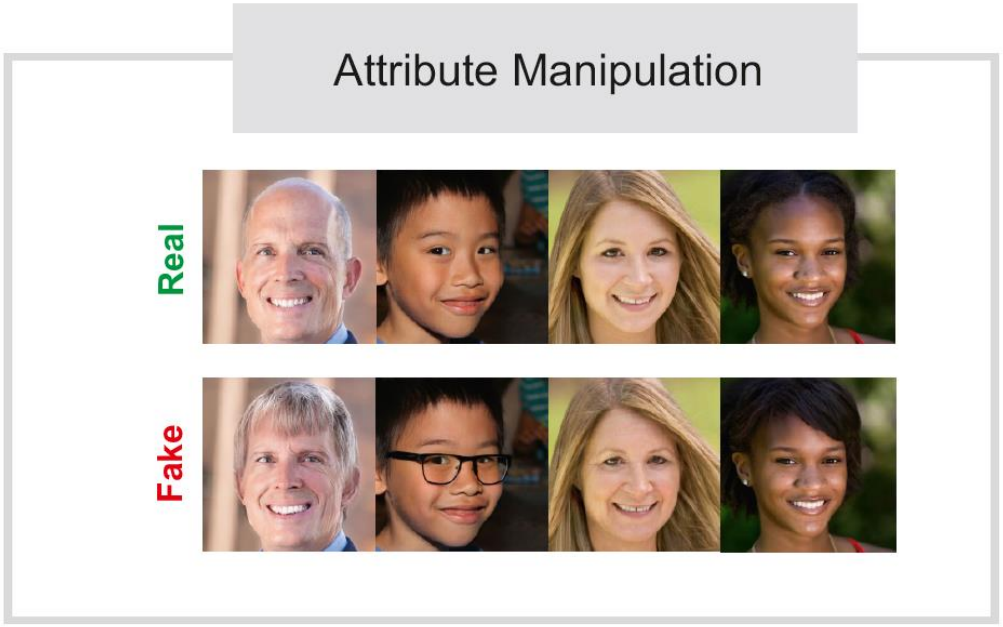

(g) Original 1 (h) Original 2 (i) LQ swap 1 → 2 (j) HQ swap 1 → 2 (k) LQ swap 2 → 1 (l) HQ swap 2 → 1

Fig. 1: Screenshot of the original videos from VidTIMIT database and low (LQ) and high quality (HQ) Deepfake videos.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, *64*, 131-148.

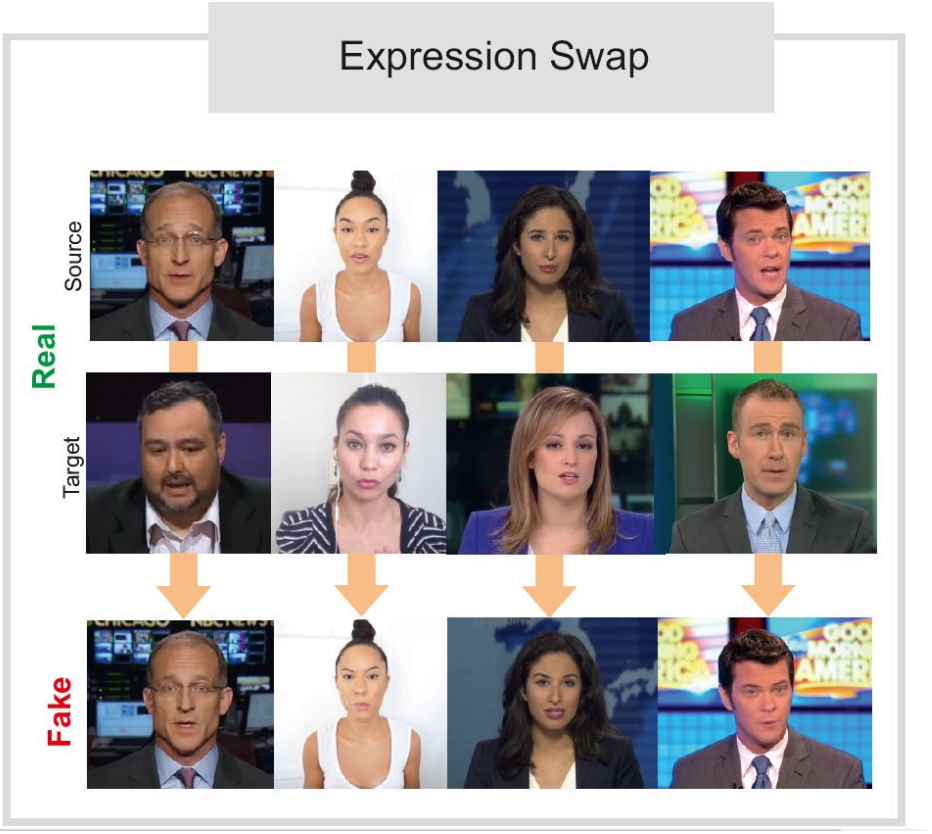
Korshunov, P., & Marcel, S. (2018). Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*.

# Attribute Manipulation




Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, *64*, 131-148.

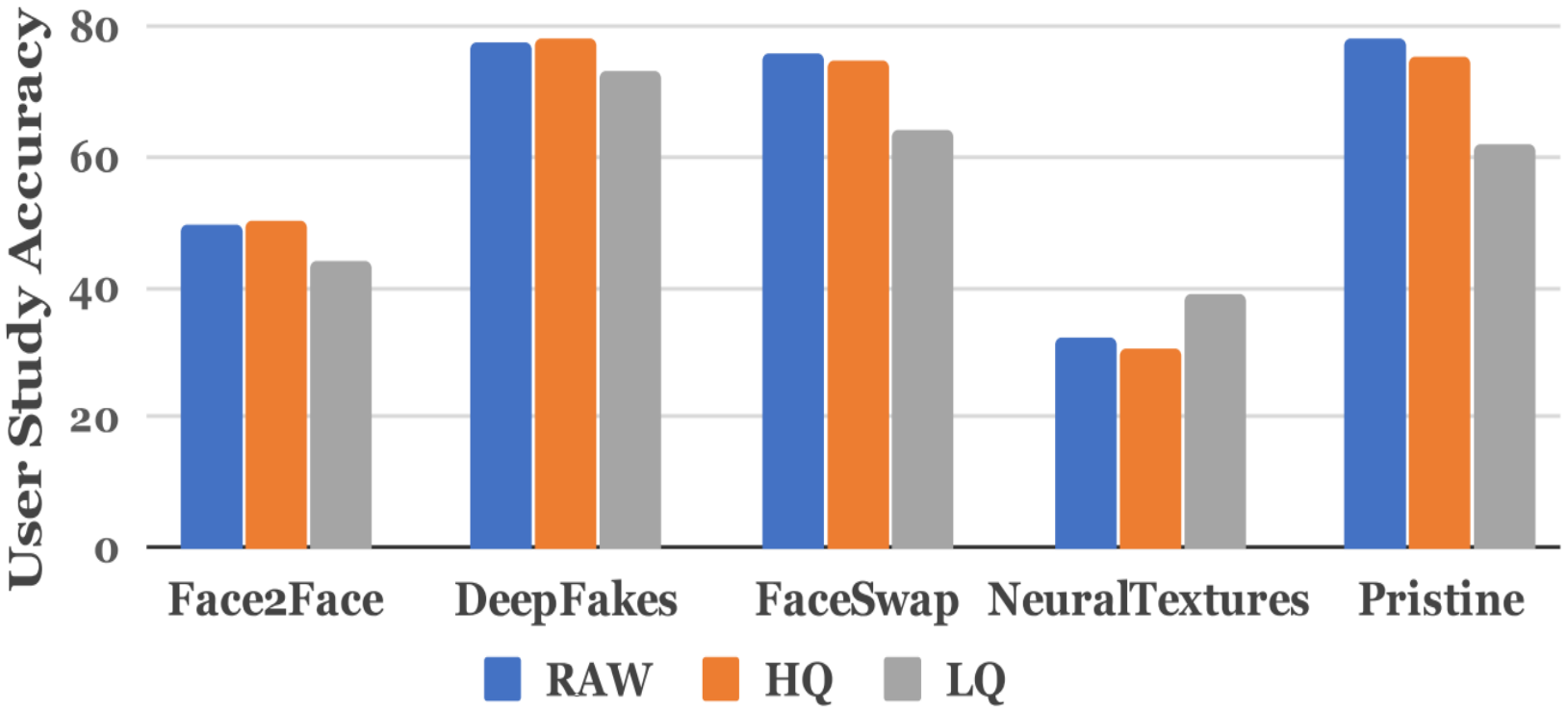
# Expression Swap




Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, *64*, 131-148.

# How to detect Deepfackes

# Humans Can't Detect Image Manipulation Well




A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics : Learning to Detect Manipulated Facial Images," arXiv, Apr. 2019.

# Deepfake Detection Methodologies

- Signal level (sensor noise, double JPEG compression, etc.)
- Physical level (lighting conditions, shadows, reflections, etc.)
- Semantic level (consistency of meta-data)

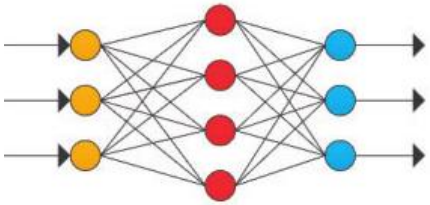
# Deepfake Detection Methodologies

- Physiological signals (breathing, pulse, eye blinking, etc.)

- Video authentication (ex. blockchain)

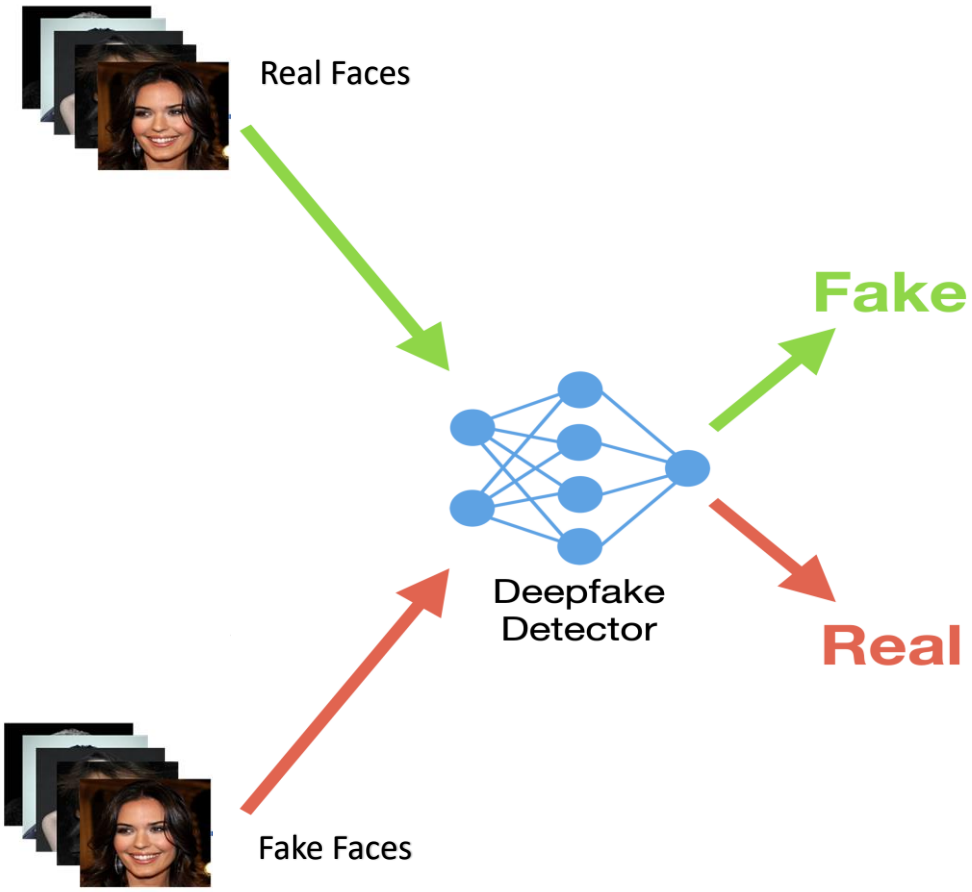
# Detection Approaches

- Detection using Machine Learning/Deep Learning



- Detection using GAN fingerprints

# Machine Learning/Deep Learning



# Dataset

cs.albany.edu/~lsw/celeb-deepfakeforensics.html

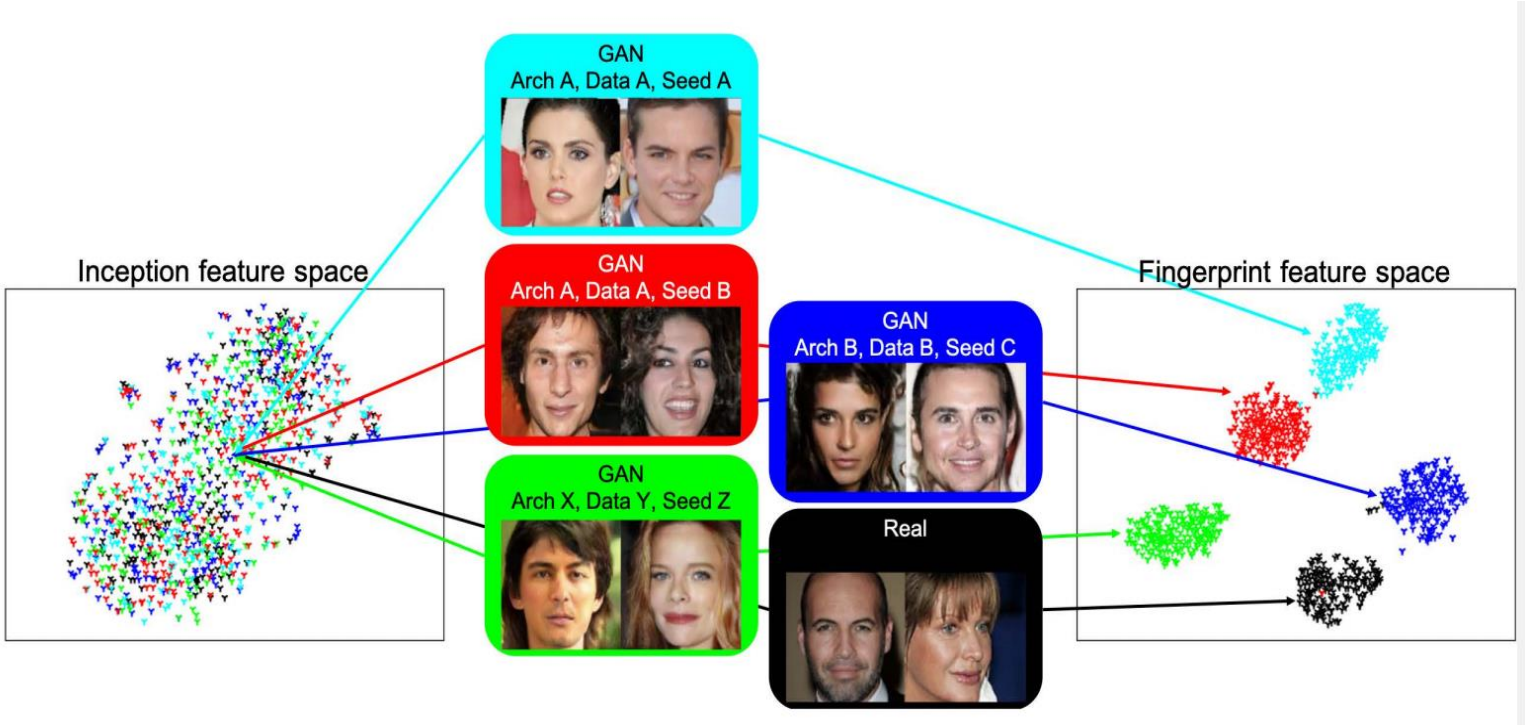
## Celeb-DF (v2): A New Dataset for DeepFake Forensics

Yuezun Li[1], Xin Yang[1], Pu Sun[2], Honggang Qi[2] and Siwei Lyu[1]
[1] University at Albany, State University of New York, USA
[2] University of Chinese Academy of Sciences, China

Github Paper Celeb-DF (v1)

Green box: Real images, Red box: Corresponding DeepFake images.

# Using GAN Fingerprint




N. Yu, L. Davis, and M. Fritz, "Attributing Fake Images to GANs: Analyzing Fingerprints in Generated Images," arXiv preprint arXiv:1811.08180, 2018.

# GAN-fingerprint removal approach

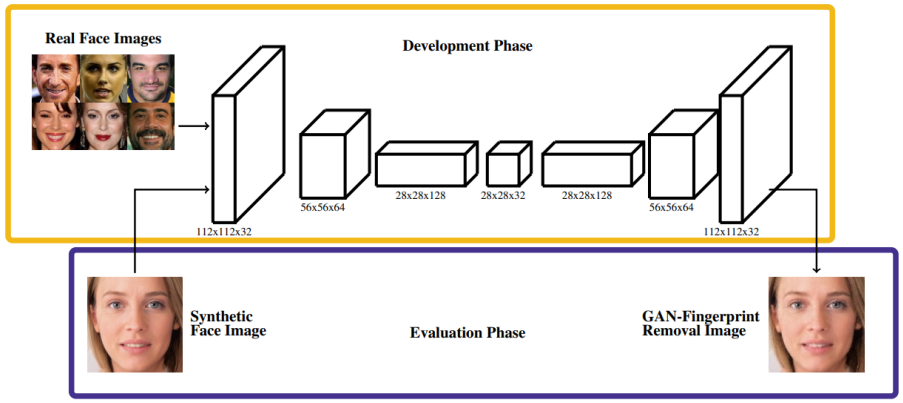
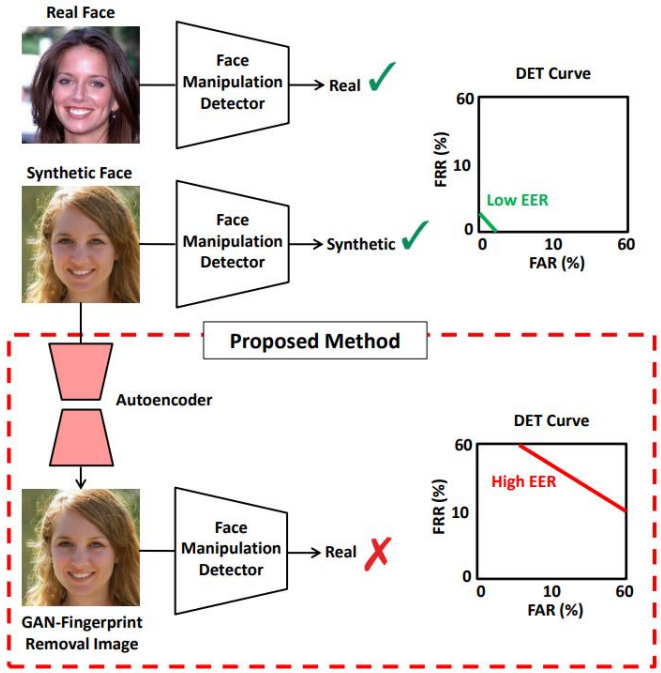



Fig. 2. **Proposed GAN-fingerprint removal approach based on convolutional autoencoders.** An autoencoder is trained with real face images from the development dataset. In the evaluation stage, once the autoencoder is trained, we can pass synthetic face images through it to remove the GAN fingerprint information, and also incorporate the correlations of real face images into the synthetic face images.




N. Yu, L. Davis, and M. Fritz, "Attributing Fake Images to GANs: Analyzing Fingerprints in Generated Images," arXiv preprint arXiv:1811.08180, 2018.

# Problem with current detection approaches

- most current face manipulations seem easy to be detected under controlled scenarios, i.e., when fake detectors are evaluated in the same conditions they are trained for.

- facial manipulation techniques are continuously improving.

These factors motivate further research on the generalization ability of the fake detectors against unseen conditions.

# The Real Business Impacts of Deepfakes

- False claims of malfeasance, damaging a product or company's reputation
- Endorsements that are not real (you thought fake written reviews were harmful)
- Video-backed HR complaints about a co-worker or a boss
- Insurance fraud, support by "video proof"
- False news about the company's owners, founders, leaders, etc.

https://www.nojitter.com/security/real-business-impacts-deepfakes

# The Real Business Impacts of Deepfakes

- Onboarding processes subverted and fraudulent accounts created
- Identity theft, using video to convince someone to alter critical personal data
- Diversion of shipments
- Orders for unwanted materials
- Payments and/or funds transfer fraudulently authorized
- Blackmail based on the threat to release a damaging vid

https://www.nojitter.com/security/real-business-impacts-deepfakes

# How to guard your organization against deepfakes

- **Employee training and awareness**

- By offering adequate training and creating awareness employees can be turned into an additional line of defense.

- Training should focus on how the technology is leveraged in malicious attempts and how this can be detected: enabling employees to spot deepfake-based social engineering attempts.

# How to guard your organization against deepfakes

- **Detection model**

- Detecting false media early can help minimize the impact on your organization.

- Developing new models that can detect fake images and videos.

# How to guard your organization against deepfakes

- **Response Strategy**

- Ensure that your organization is ready to adequately respond to a deepfake.

- Have a plan in place that can be set in motion when a deepfake is detected.

- It's important that individual responsibilities and required actions are defined in this plan.