DEEP LEARNING FOR DEEPFAKE CREATION AND DETECTION

Aman Kumar Singh
Student, IIT Kanpur
amanks20@iitk.ac.in

Dr. Nishchal K. Verma
Professor, Dept. of Electrical Engineering, IIT Kanpur
nishchal@iitk.ac.in

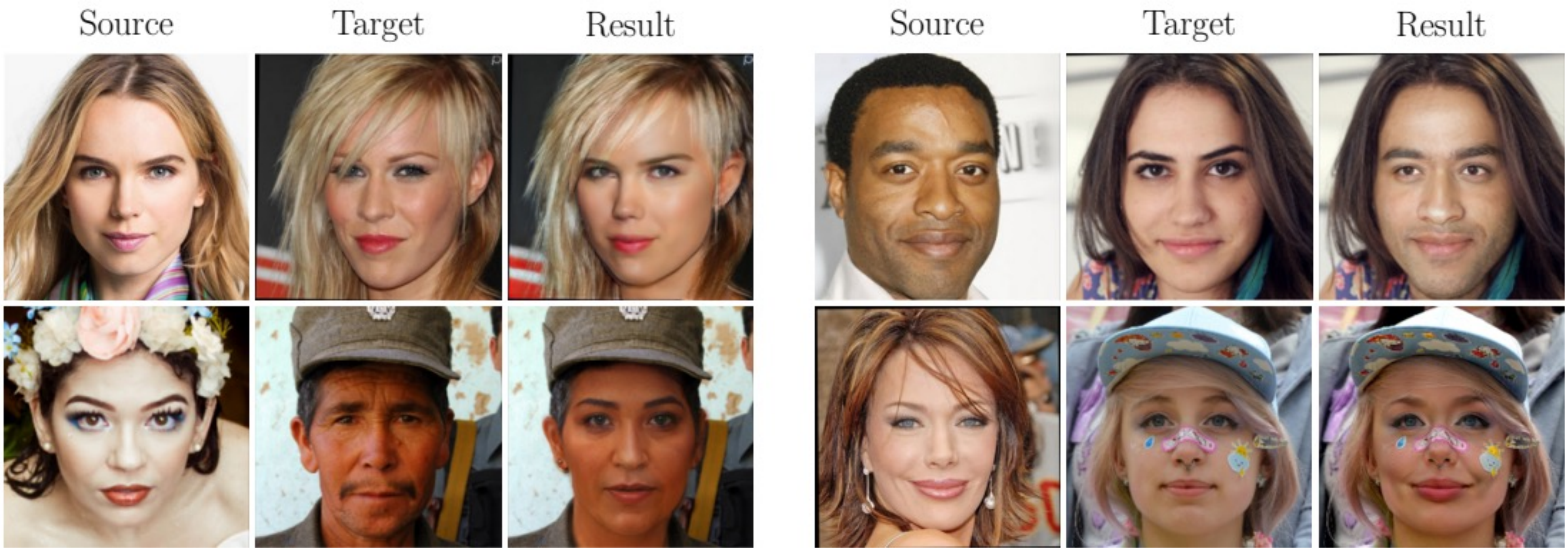
Introduction to Deepfakes

- DEEPFAKE = **DEEP** (Deep Learning) + **FAKE**
- In a narrow definition, Deepfakes involve techniques that superimpose face images of a target person onto a video of a source person to create a video where the target person appears to do or say things that the source person does
- In a broader definition, Deepfakes are AI-synthesized content creating realistic but fake visual media
- Categories: ***Face-swap***, ***lip-sync***, ***puppet-master***
- Created using deep learning models like ***Autoencoders*** and ***generative adversarial networks (GANs)***
- Various detection methods, mainly based on deep learning, aim to identify inconsistencies in facial features, temporal discrepancies, and visual artifacts within manipulated content
- Detecting deepfakes is challenging due to their increasingly convincing nature

Introduction to Deepfakes

▶ *Face-swap*

Superimpose the face of a source person onto an image/video of a target person

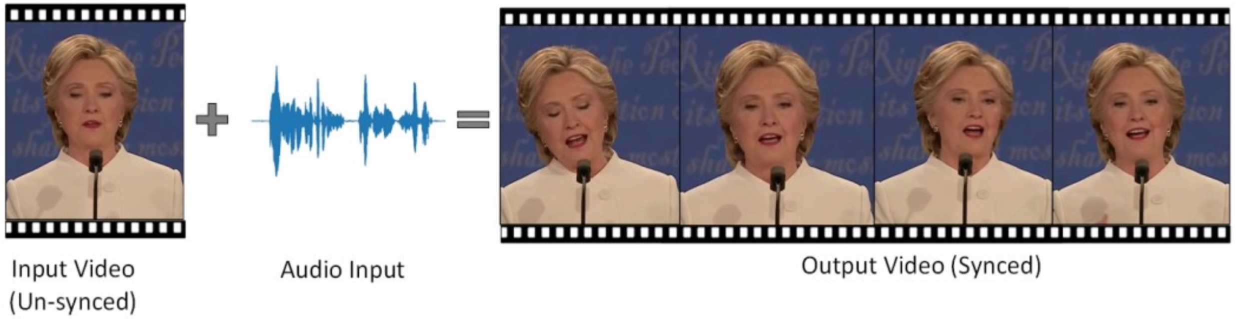


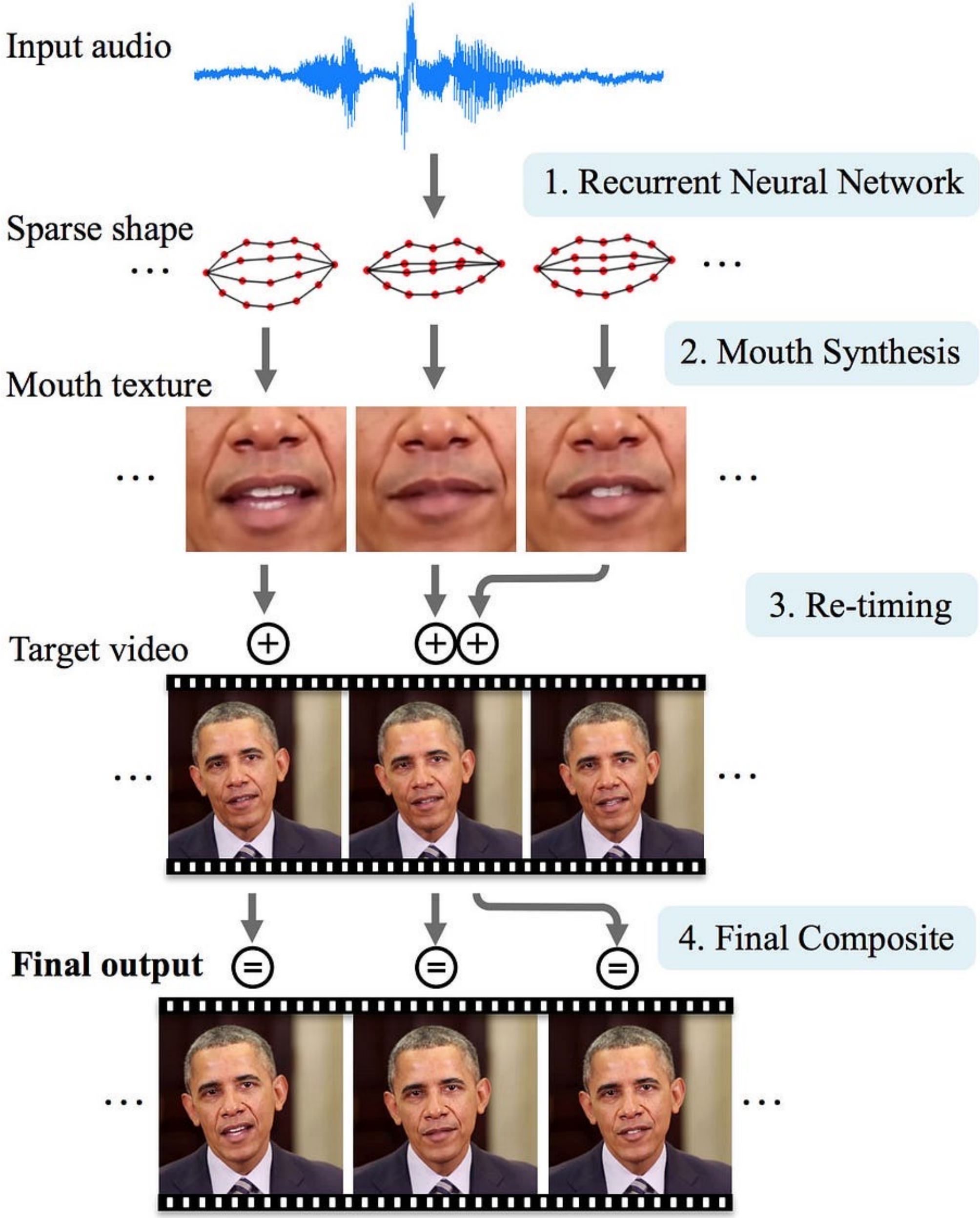
Source: Yang, X. & Bo, H. High-Fidelity Face Swapping with Style Blending. (2023).

Introduction to Deepfakes

▶ *Lip-sync*

Make the mouth movements consistent with an audio recording



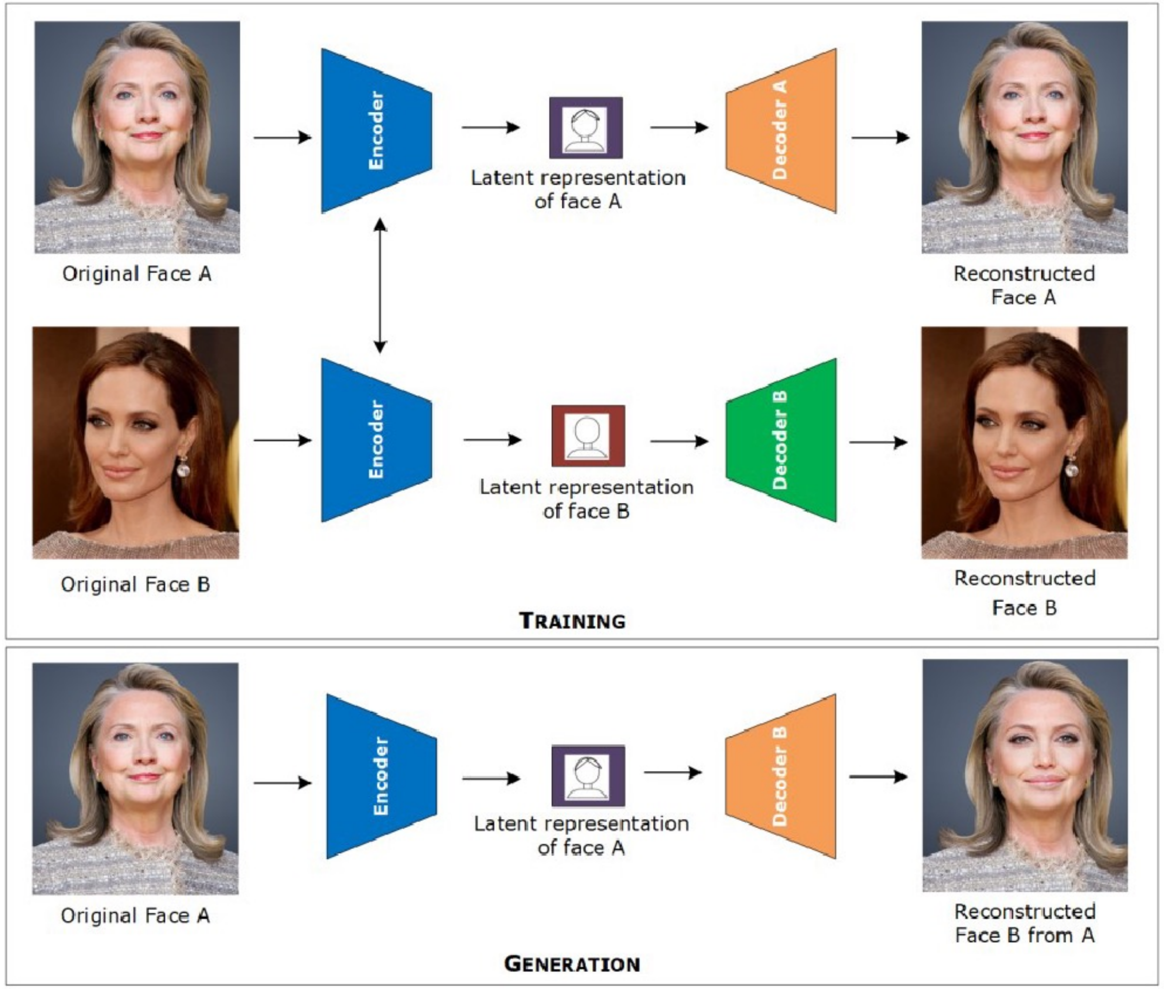


Source: Masood, M. *et al. (2023)*. and Lip-Syncing Obama at https://www.washington.edu/news/2017/07/11/lip-syncing-obama-new-tools-turn-audio-clips-into-realistic-video/

Introduction to Deepfakes

▶ *Puppet-master*

Make the target person mimic facial expressions, eye, and head movements of a source person in real-time



Source: Masood, M. *et al*. Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. *Appl Intell* **53**, 3974–4026 (2023).

Significance and Impact of Deepfakes

Positive Significance:

- Restoring lost voices and recreating historical figures
- Enhancing artistic expression in comedy, cinema, music, and gaming
- Aid individuals with disabilities in expressing themselves online
- Improving medical training with realistic images and scenarios

Negative Implications:

- Spreading propaganda and fake news
- Influencing elections and public opinion
- Damaging the reputation of public figures
- Creating non-consensual pornography
- Eroding trust in institutions and media
- Violating privacy and harming mental health






Source: Images generated with AI, Microsoft Copilot Designer, *Bing Image Creator (2024)*.

Significance and Impact of Deepfakes

- However, no. of malicious uses of deepfakes largely dominates that of positive ones
- Advanced deep neural networks and abundant data have made forged content nearly indistinguishable from authentic ones
- Creation of manipulated content is simplified, requiring only a target individual's identity photo or a short video
- Little effort yields highly convincing fake footage, even from still images.
- Deepfakes pose a threat to both public figures and ordinary individuals, evidenced by scams via voice deepfakes
- DeepNude software poses more disturbing threats as it can create non-consensual porn of any person
- Chinese app Zao allows less-skilled users to superimpose their faces onto movie stars', potentially misused in movies and TV clips

Deepfake Creation

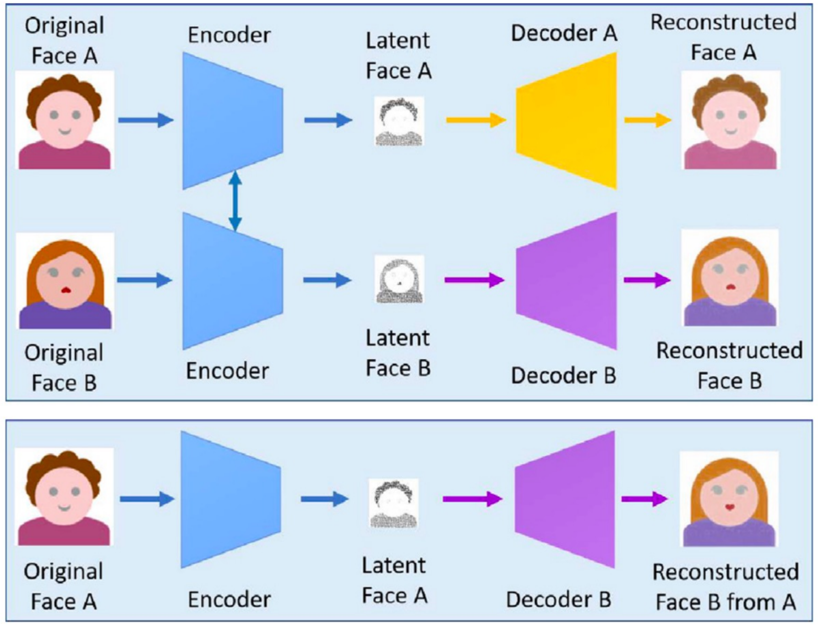
- Deepfakes are created using deep learning models, particularly ***Autoencoders*** and ***generative adversarial networks (GANs)***
- Autoencoders encode facial features from source images and decode them onto target images, forming the basis for early deepfake creation tools like FakeApp
- GANs use generator against a discriminator network in a minmax game to produce convincing fake images (eg: faceswap-GAN and StyleGAN use advanced GAN to generate highly realistic deepfakes)
- Adding loss functions like adversarial and perceptual losses in GAN models enhance the quality and realism of deepfakes, making them harder to detect
- Target individuals are mostly celebrities, politicians, public figures & sometimes ordinary people

Images generated with AI (2023).

Autoencoders

- An autoencoder is a neural network with two functions: encoding & decoding
- **Encoding** to extract facial features from source images and compress into lower dimensional latent face
- **Decoding** to reconstruct the faces from latent face representation
- In deepfake creation, autoencoders facilitate face swapping by connecting features from one face to the decoder of another face
- First deepfake creation attempt was FakeApp, utilizing autoencoder-decoder pairing to swap face between source and target images
- Tools like DeepFaceLab, DFaker, and DeepFake_tf employ this strategy

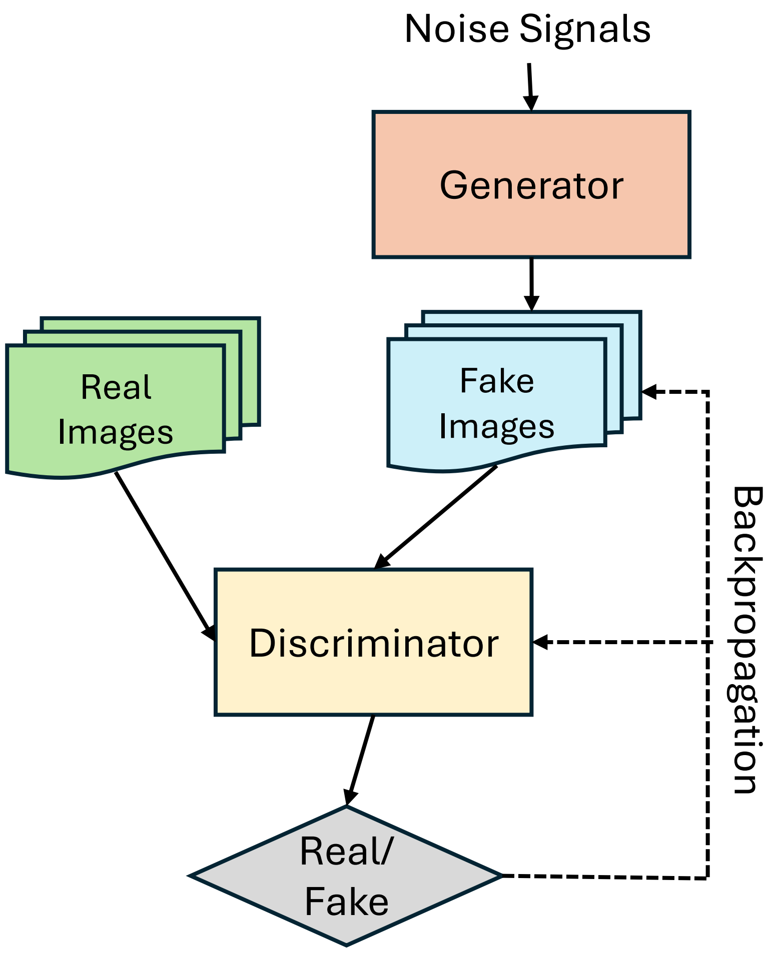
Source: Yang, X. & Bo, H. (2023).



Source: Nguyen, T. T. *et al.* (2022).

GANs

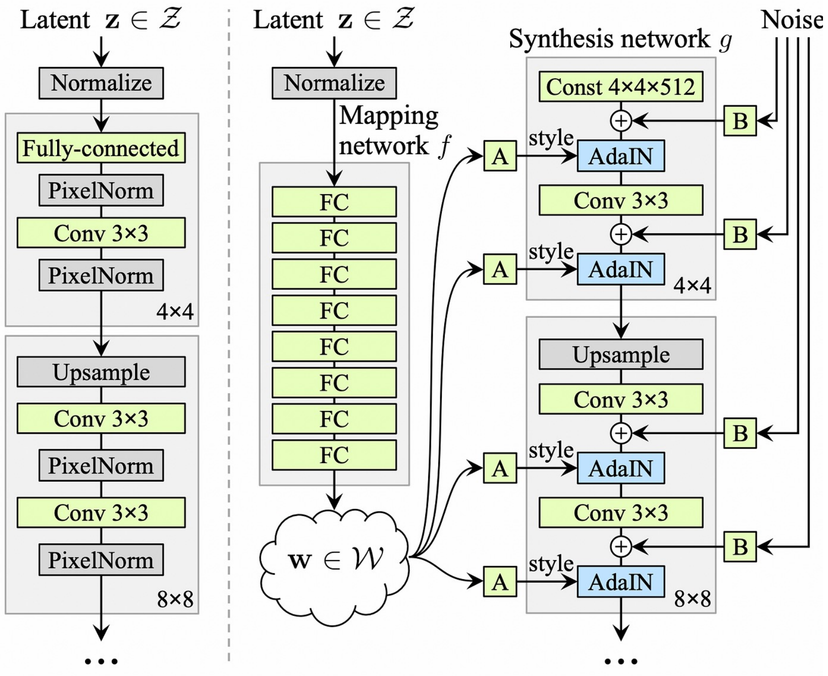
- Generative Adversarial Networks (GANs) are a type of AI framework with two neural networks: a generator and a discriminator
- **Generator** creates synthetic data (e.g., images) from random noise
- **Discriminator** acts as a critic to distinguish real and fake data
- GANs operate through **adversarial training**, where the generator aims to create increasingly realistic data to fool the discriminator, while the discriminator strives to become better at distinguishing real from fake data.
- GAN training is akin to a minimax game, where the generator's goal is to produce indistinguishable data from real data, while the discriminator aims to correctly classify real and fake data



Source: Nguyen, T. T. *et al.* (2022).

StyleGAN

- StyleGAN is an advanced GAN architecture developed by NVIDEA to create high-quality and realistic images, particularly of human faces
- Unlike traditional GAN, StyleGAN incorporates a mapping network, synthesis network and employs two latent codes during training
- StyleGAN integrates advanced techniques such as adaptive instance normalization (AdaIN) operations to improve realism and generate visually indistinguishable images
- By manipulating latent codes and styles, users can produce highly convincing deepfake content with nuanced facial expressions and features



Source: Karras, T., Laine, S. & Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. (2019).

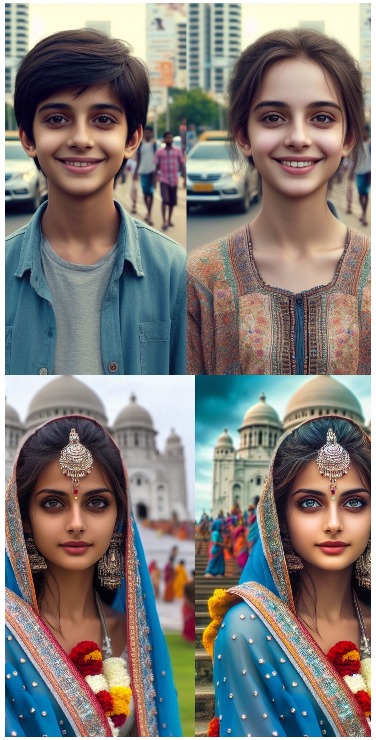
StyleGAN

▶ Images generated by StyleGAN

The individuals depicted in these images do not exist but are generated by artificial intelligence through the analysis of portraits.
Source: Karras, T., Laine, S. & Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. (2019).

Enhancements for Deepfake

- Loss functions such as **adversarial** and **perceptual losses** are incorporated into GANs like faceswap-GAN to enhance realism of deepfakes and address issues like artifacts, eye movements, etc
- Adversarial loss function focuses on training the generator to produce realistic images by fooling the discriminator
- Perceptual loss function concentrates on ensuring that the generated images match high-level features of real images, enhancing realism



Images generated by AI *(2024)*.

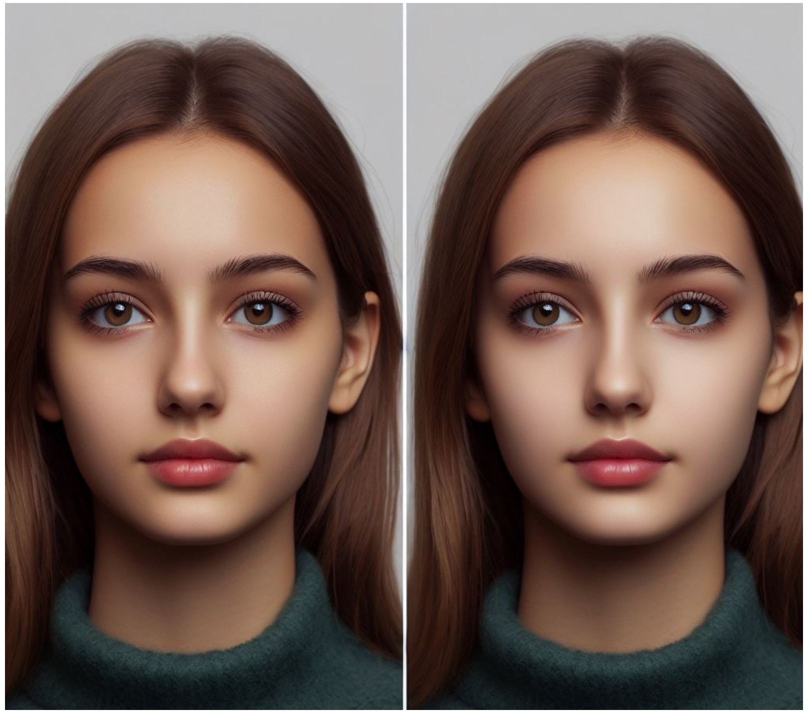
Loss Function	Adversarial Loss	Perceptual Loss
Objective	Enhance overall realism of images	Improve perceptual similarity to originals
Calculation	Based on probability distribution	Based on visual appearance similarity
Optimization	Maximizes realism	Minimizes perceptual difference
Training Behaviour	Emphasizes overall image quality	Focuses on preserving original feature
Evaluation Metric	Discriminator's classification	Feature-based comparison with originals

Deepfake Detection

- Deepfake detection is a **binary classification problem** involving distinguishing between authentic and tampered videos
- Deepfake detection is done using **classifiers** trained on extensive datasets containing both real and fake images/videos
- Urgent development is needed for robust methods to detect deepfakes from genuine images and videos

Source: Karras, T., Laine, S. & Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. (2019).

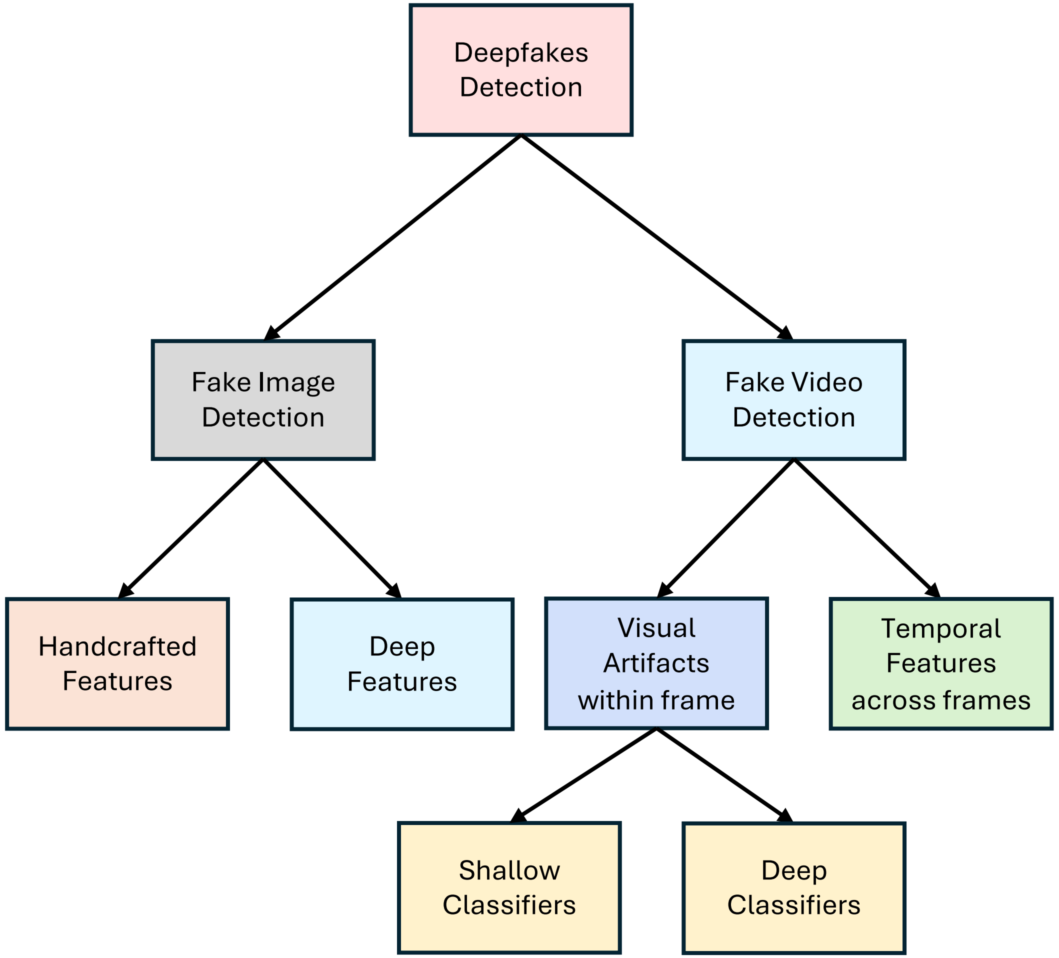
Deepfake Real



The young woman depicted in the image do not exist but is generated by AI just for visualisation. Bing Image Creator (2024).

Deepfake Detection

- Handcrafted features are attributes manually designed and extracted from data
- Deep Features are high-level representations of data automatically learned by deep neural networks through training
- Visual Artifacts are anomalies or inconsistencies within a single frame of an image or video that indicate manipulation or alteration that may not always be obvious or visible to the naked eye
- Temporal features span multiple frames of a video sequence, capturing changes and patterns over time

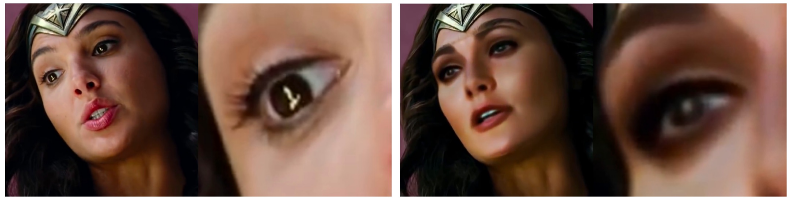


Fake Image Detection

- **Handcrafted Features:** Early methods relied on manually extracted features highlighting inconsistencies in fake image synthesis
- **Deep Features:** Recent advancements leverage deep learning to automatically extract discriminative deep features for more robust detection like Lima et al. (2020) and Amerini and Caldelli (2020)

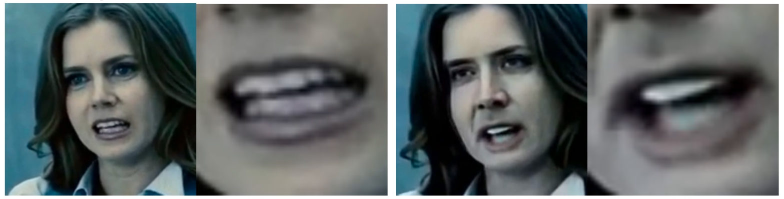
Difference between left and right eye color

Source: Matern, F., Riess, C. & Stamminger, M. Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations (2019).



Missing reflection details in eyes

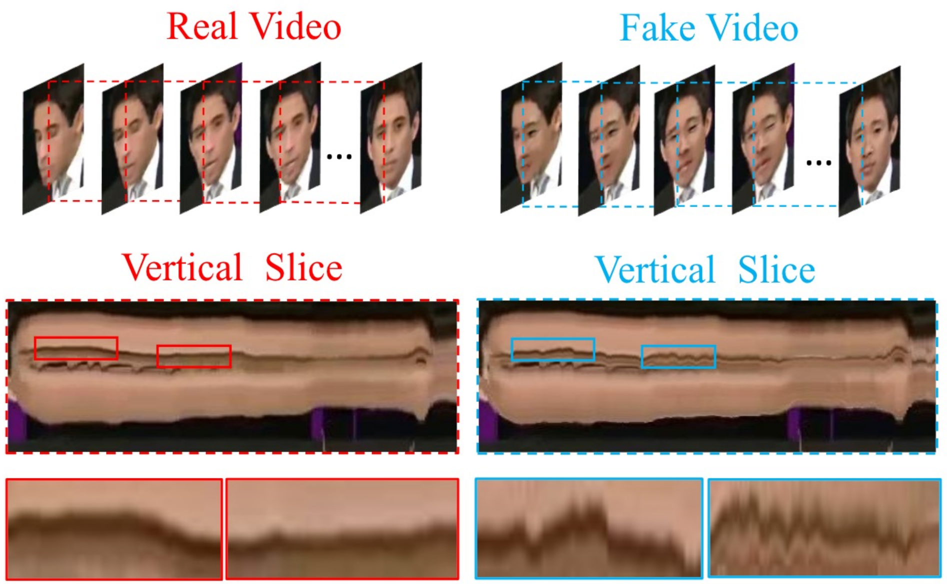
Shading artifacts



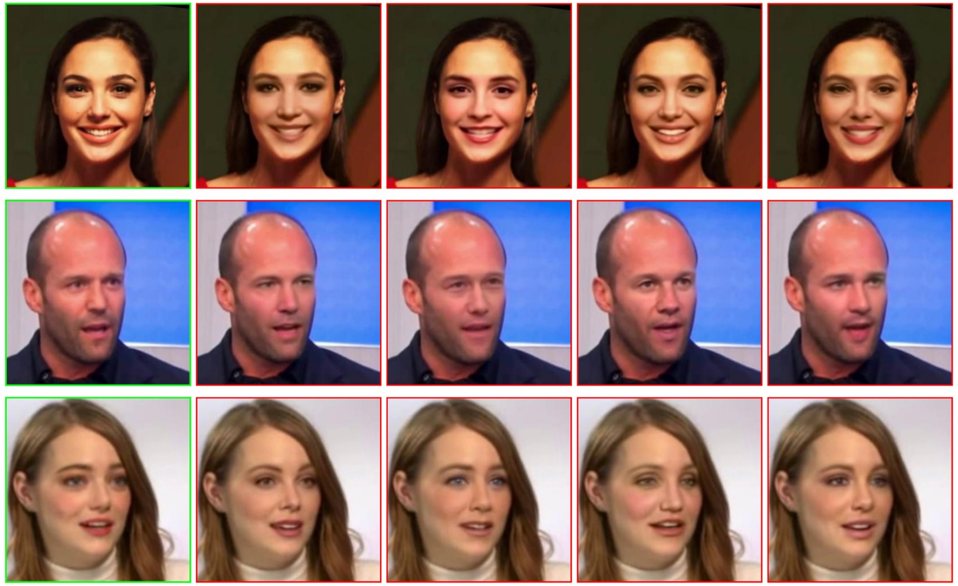
Missing geometry (teeth as structureless white blob)

Fake Video Detection

- **Temporal Features across frames:** Exploit temporal feature inconsistencies across video frames to detect deepfakes
- **Visual Artifacts within frame:** Identify visual artifacts or anomalies within single frames to differentiate between real and fake videos (shallow and deep classifiers)



Source: Gu, Z. *et al.* (2021).



Source: Khormali, A. & Yuan, J.-S. (2021).

Challenges and Future Directions

Challenges

- Deepfakes pose a growing threat, especially in politics and security
- Detecting deepfakes is struggling to keep up with their increasing quality
- Limited data hampers the ability of detection models to adapt
- Models often can't handle real-world scenarios or unknown variations
- Attacks on detection systems are exploiting vulnerabilities

Future Directions

- People need tools to think critically and fight misinformation
- Advancing AI for real-time detection to stop sophisticated deepfake attempts
- Making easy-to-use tools for everyone to verify media
- Creating transparent evidence in legal cases using AI

Critical Analysis

- Combining different modalities such as visual, audio, and linguistic cues can enhance the accuracy of deepfake detection to detect deepfake by analysing discrepancies between them
- Deepfakes often lack the subtle, involuntary movements characteristic of genuine human expressions, watching for unnatural behaviors like micro-expressions can help detect deepfakes
- Leveraging blockchain technology to establish and verify the authenticity of media content can help detect deepfake
- Continuously pitting detection models against advanced generative models in adversarial settings can bolster resilience

References

[1] Nguyen, Thanh Thi *et al.* (2022) 'Deep learning for deepfakes creation and detection: A survey', *Computer Vision and Image Understanding*, 223, p. 103525. Available at: https://doi.org/10.1016/j.cviu.2022.103525.

[2] Yang, X. and Bo, H. (2023) 'High-Fidelity Face Swapping with Style Blending'. arXiv. Available at: http://arxiv.org/abs/2312.10843 (Accessed: 23 March 2024).

[3] Karras, T., Laine, S. & Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. Preprint at http://arxiv.org/abs/1812.04948 (2019).

[4] Masood, M. *et al.* Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. *Appl Intell* **53**, 3974–4026 (2023).

[5] Gu, Z. *et al.* Spatiotemporal Inconsistency Learning for DeepFake Video Detection. Preprint at http://arxiv.org/abs/2109.01860 (2021).

[6] Matern, F., Riess, C. & Stamminger, M. Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations. in *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)* 83–92 (IEEE, Waikoloa Village, HI, USA, 2019). doi:10.1109/WACVW.2019.00020.

[7] Khormali, A. & Yuan, J.-S. ADD: Attention-Based DeepFake Detection Approach. *BDCC* **5**, 49 (2021).

[8] Gambín, Á. F. Deepfakes: current and future trends.

[9] Heidari, A., Jafari Navimipour, N., Dag, H. & Unal, M. Deepfake detection using deep learning methods: A systematic and comprehensive review. *WIREs Data Min & Knowl* **14**, e1520 (2024).

[10] Microsoft Copilot Designer, 'Bing Image Creator' (2024). All images, unless otherwise cited, were generated using AI.

Thank You!

Aman Kumar Singh (208070100)
amanks20@iitk.ac.in