INT307 Multimedia Security System

Deep Fake Detection

Shengchen Li

Xi'an Jiaotong-Liverpool University

20th Nov 2022

Aims

- Understand basic principles of fake face generation
- Know the basic way to attack a face recognition system

Deepfake

- Generation of media with deep learning techniques
- Manipulation of media to mislead public / other parties

Definition

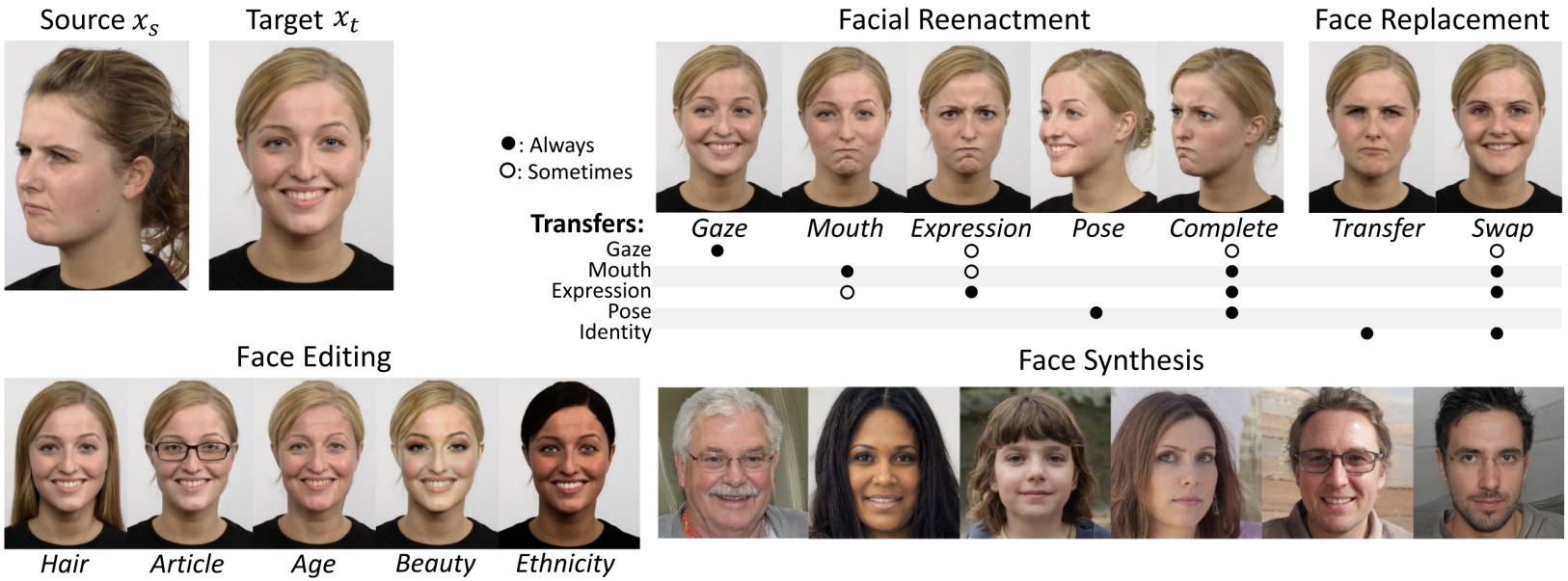
Believable media generated by a deep neural network

Ways of Attack

- Reenactment
- Replacement
- Editing
- Synthesis

Xi'an Jiaotong-Liverpool University 西交利物浦大學

Deepfake related to human face



Reenactment

- Expression
- Mouth
- Gaze
- Pose
- Body

The Attack Model

- Defamation
- Cause Discredability
- Spread Misinformation
- Tamper with Evidence

Replacement

- Transfer
- Swap

The Attack Model

- Humiliate, Defame and Blackmail
- Disseminating Public Opinions

Editing and Synthesis

- Editing (Attributes)
  - Clothes
  - Facial Hair
  - Age
  - Weight
  - Beauty
  - Ethnicity
- Synthesis
  - No Basic Objects

Basic Architectures

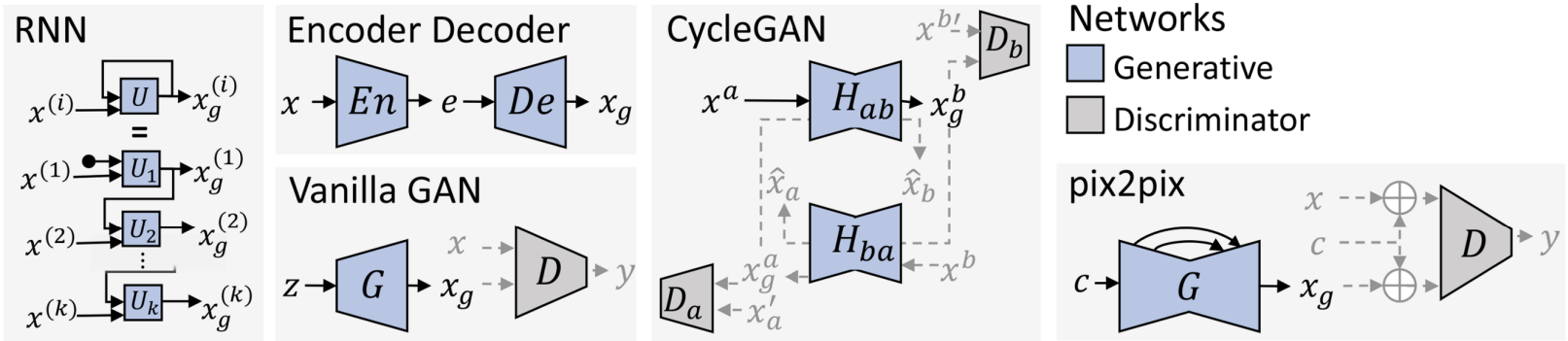


Fig. 4. Five basic neural network architectures used to create deepfakes. The lines indicate dataflows used during deployment (black) and training (gray).

Architectures of Deep Learning

- Encoder-Decoder Networks
- Convolutional Neural Network
- Generative Adversarial Networks
  - Image-to-Image Translation (pix2pix)
  - CycleGAN
- Recurrent Neural Network

Deepfake Creation Basics

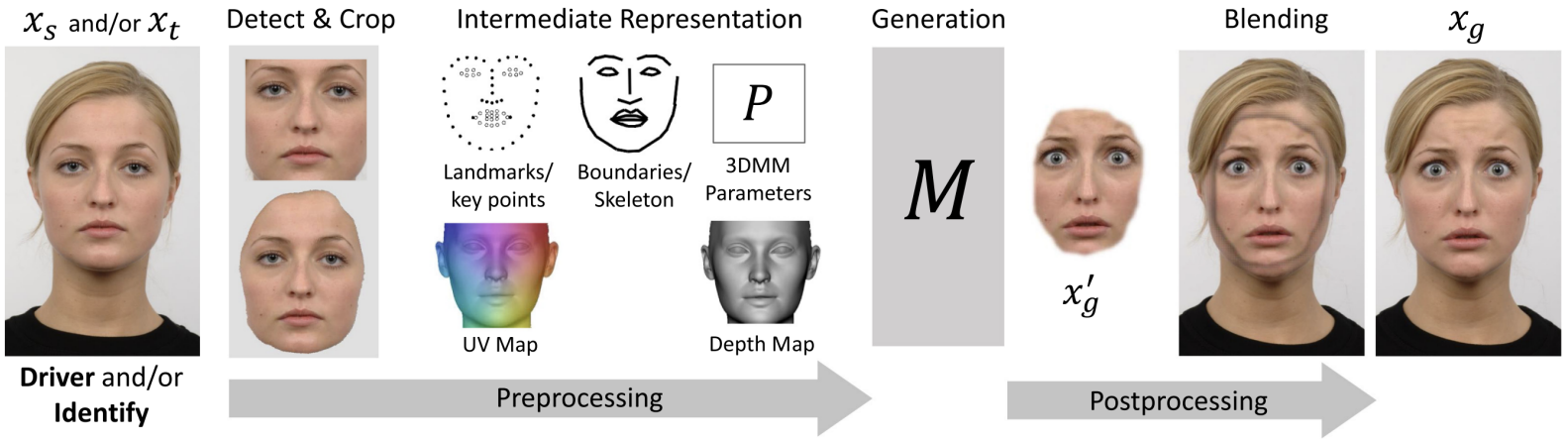


Fig. 5. The processing pipeline for making reenactment and face swap deepfakes. Usually only a subset of these steps are performed.

1. Detects and Crops the Face
2. Extracts Intermediate Representation
3. Generates a New Face based on Driving Signal
4. Blend Generated Face back into the Target Frame

Approaches to Drive an Image

- Direct Mapping
- Feature Disentanglement
- Additional Encoding
- Convert Intermediate Face / Body Representation
- Use the Optical Flow Field
- Create a Composite of the Original Content

Challenges

- Generalisation
- Paired Training
- Identity Leakage
- Occlusions
- Temporal Coherence

Counter Measures

Artifact-Specific Detection

- Blending
- Environment
- Forensics
- Behaviour
- Physiology
- Synchronisation
- Coherence

Counter Measures

Undirected Approaches

- Classification
- Anomaly Detection

Discussion

What are the advantages and disadvantages for both methods?

The Creation of Deepfakes

Trade-off between Methodologies

- Data vs Quality
- Speed vs Quality
- Availability vs Quality

Research Trends

- Unpaired Self-Supervised Training Techniques
- One / Few-shot learning
- Real-time

Discussion

- How can you attack a face recognition system?