Unmasking Deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence

Fakhar Abbas[1], and Araz Taeihagh[1,2]

1 Centre for Trusted Internet and Community, National University of Singapore

2 Lee Kuan Yew School of Public Policy, National University of Singapore

Paper QR Code

Introduction

Focus: The study systematically reviews current techniques for detecting and generating deepfakes, focusing on frameworks, algorithms, tools, and practical challenges. We highlight the growing concern over the impact of deepfakes on politics, media, and society and propose policy recommendations.

Proposed Solution We analyze both the detection and generation aspects of deepfakes, addressing their ethical implications and offering policy solutions to safeguard digital integrity.

Key objectives

- Investigate both the harmful and creative aspects of deepfake technology.
- Explore current AI-based methods for deepfake detection (e.g., face-swapping, face reenactment) and generation (e.g., GANs, diffusion models).
- Highlight limitations in the generalization, accuracy, and real-world application of detection methods.
- Discuss privacy concerns, disinformation, and the erosion of trust in digital media caused by deepfakes.
- Propose actionable policy measures to mitigate the impact of deepfakes, focusing on ethical AI governance, platform accountability, and legal frameworks

Research Questions (RQs)

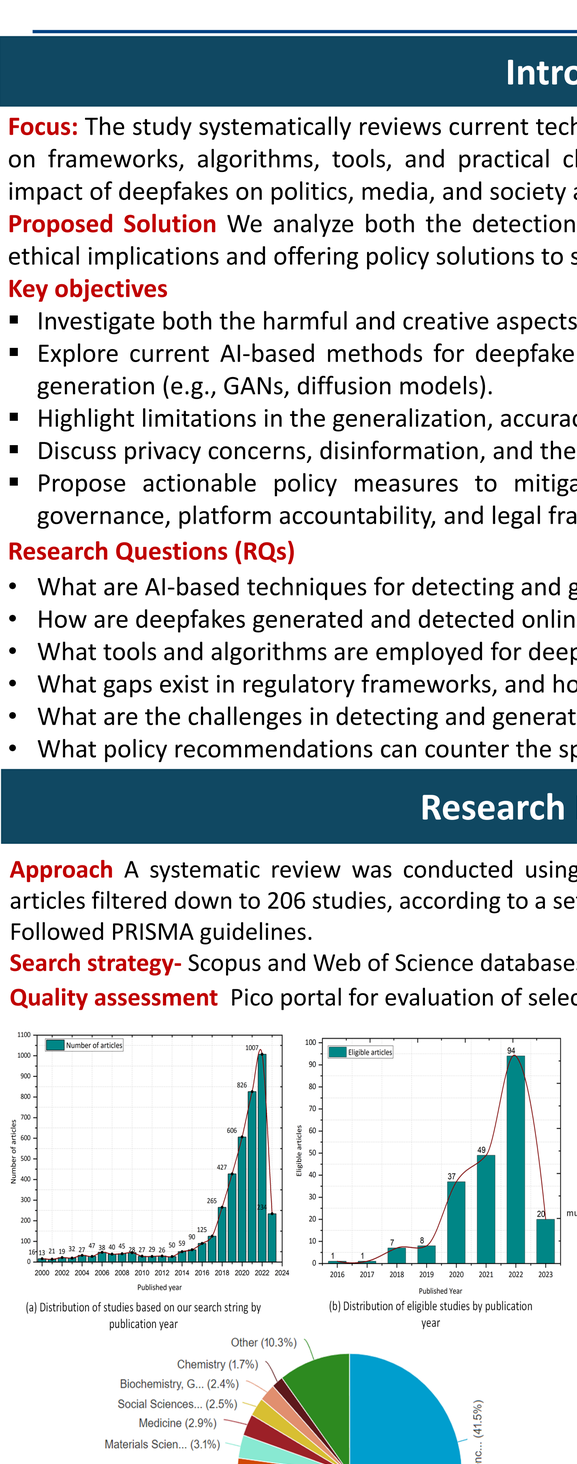
- What are AI-based techniques for detecting and generating deepfakes?
- How are deepfakes generated and detected online using AI?
- What tools and algorithms are employed for deepfake detection and generation?
- What gaps exist in regulatory frameworks, and how can global harmonization benefit society?
- What are the challenges in detecting and generating deepfakes?
- What policy recommendations can counter the spread of deepfakes?

Research Methodology

Approach A systematic review was conducted using Scopus and Web of Science databases, with over 4,000 articles filtered down to 206 studies, according to a set of inclusion and exclusion criteria.
Followed PRISMA guidelines.

Search strategy- Scopus and Web of Science databases

Quality assessment Pico portal for evaluation of selected studies



(a) Distribution of studies based on our search string by publication year

(b) Distribution of eligible studies by publication year

(c) Distribution of studies based on our search string by research area

Fig. 1 Number of publications on deepfake detection and generation, by publication year and research area

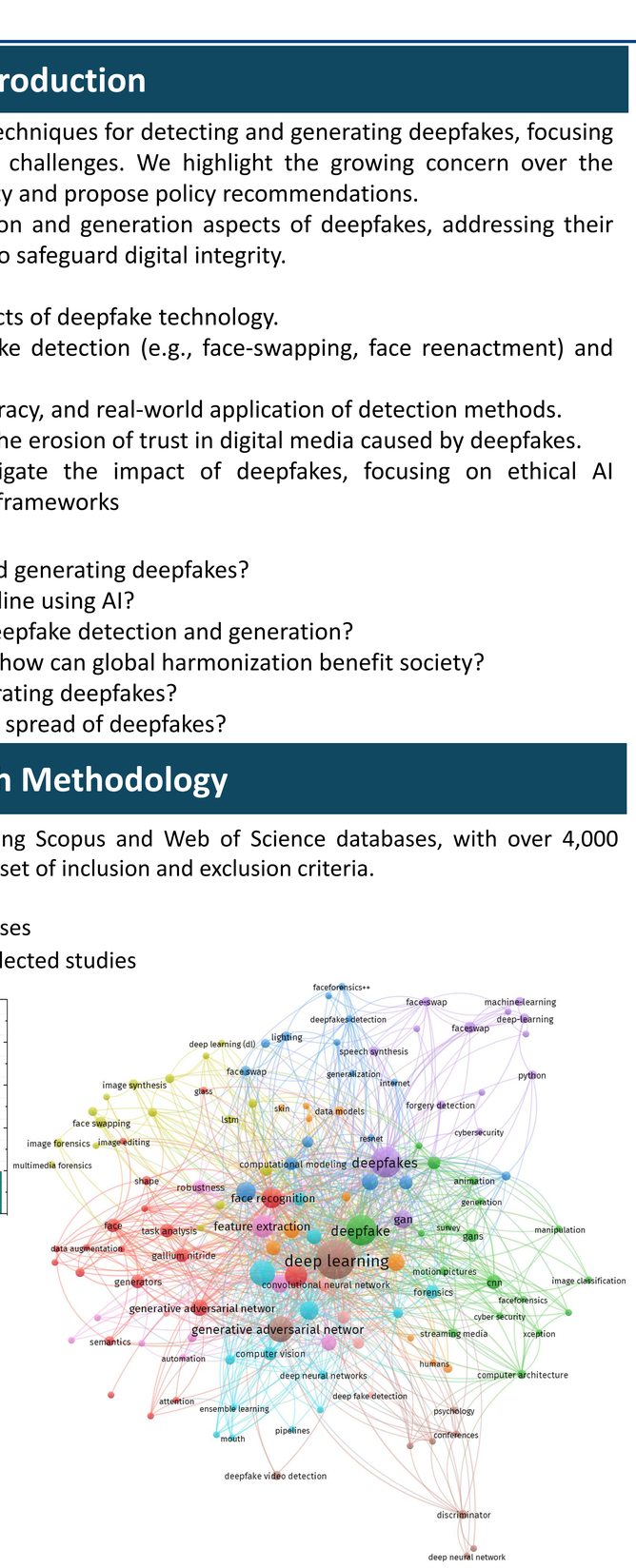


Fig. 2 Keyword co-occurrences visualization analysis of eligible studies using the VOSviewer tool

Results

A PRISMA flowchart was used for the details of the evidence

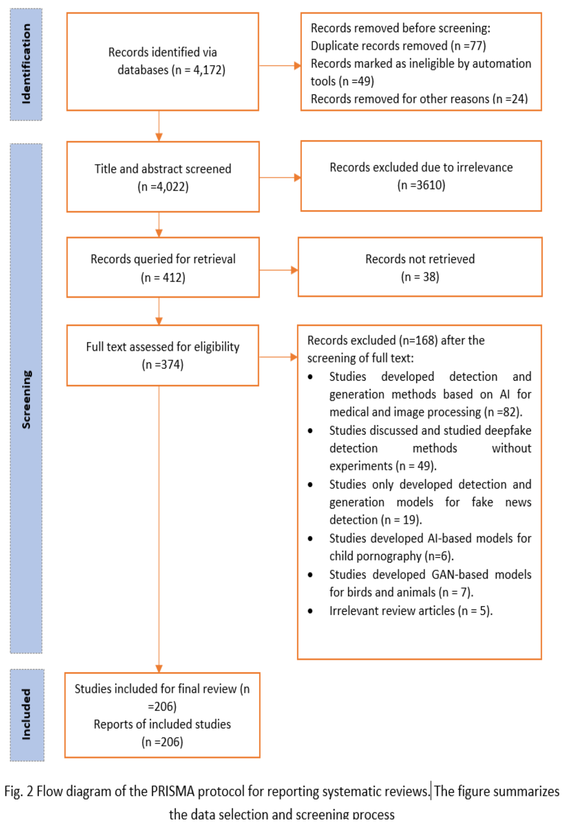


Fig. 2 Flow diagram of the PRISMA protocol for reporting systematic reviews. The figure summarizes the data selection and screening process

Deepfake detection and generation using AI

Taxonomy of deepfake detection and generation approaches.

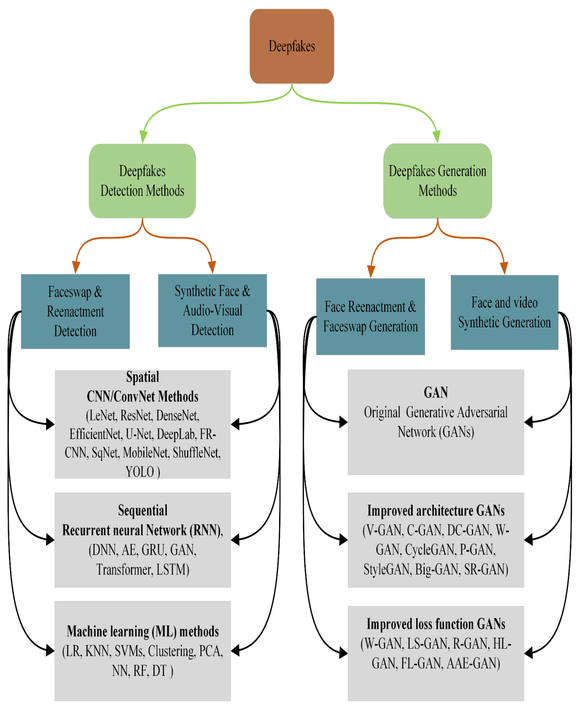


Fig. 4 Taxonomy of deepfake detection and generation approaches

Fig. 5 Face swap, reenactment, and expression manipulation, visual illustration

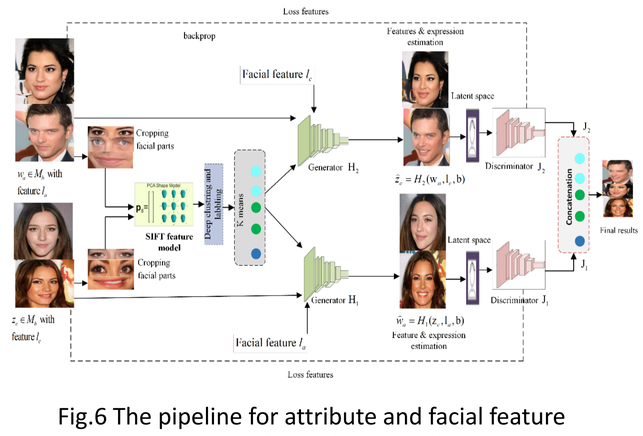


Fig.6 The pipeline for attribute and facial feature manipulation (IC-DGAN)

Deepfake Detection Methods

- Face Swap Detection: AI-based systems (e.g., CNNs, ViXNet) focus on spatial-temporal inconsistencies to detect manipulations.
- Audio-Visual Inconsistencies: AI tools analyze audio and visual cues discrepancies, revealing deepfake content (e.g., AVFakeNet).

Ethical & Legal Considerations

- **Privacy Risks:** Deepfakes can harm individual privacy by impersonating people and creating misleading content.
- **Trust in Media:** The rise of deepfakes threatens the authenticity of visual media, undermining public trust.
- **Identity Theft & Misinformation:** Malicious use of deepfakes in spreading disinformation or damaging reputations.

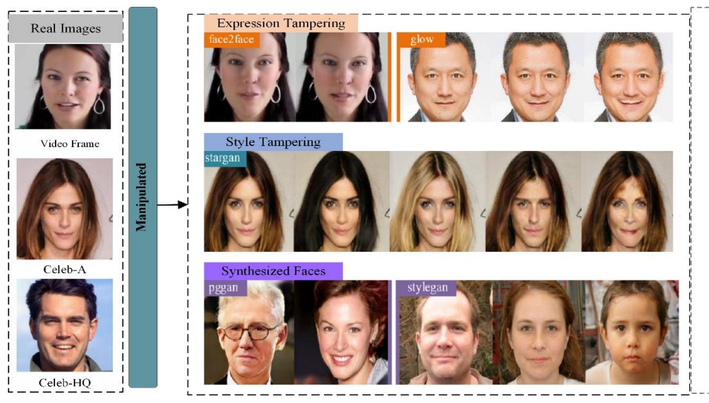


Fig. 5 Real and synthesized faces taken from the hybrid fake faces datasets

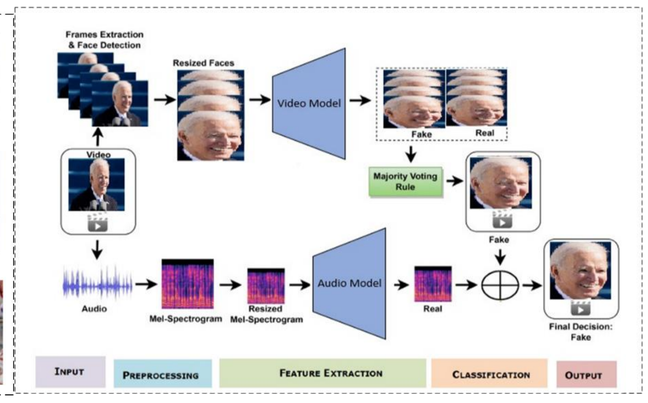


Fig. 6 Pre-processing pipeline for audio-visual deepfake detection

Key Tools for Deepfake Generation and Detection

- **StyleGAN ADA:** Adaptive augmentations stabilize high-quality image/video synthesis.
- **DiscoFaceGAN:** 3D learning framework for lifelike synthetic face generation.
- **DeepFaceLive**: Real-time face animation via 3D models and deep learning.
- **FALdetector:** CNN-based tool for precise detection of manipulated facial images.
- **Face2Face:** 3D morphable models enable real-time face reenactment in videos.
- **StarGANv2:** Multi-domain image and voice conversion via advanced GAN frameworks.
- **TAFIM:** Fusion-based method to block artificial facial manipulations in media.
- **FakeLocator:** Transformer-based model localizes manipulated regions in deepfake images.
- **DFDNet:** A high-fidelity face restoration network enhances detection by recovering details.

Policy Recommendations

Standardization & Bias Mitigation

- Develop standardized detection tools and diverse datasets to improve generalization and reduce biases.
- Ensure detection methods are not only accurate but also adaptable to unseen data.

Identity Verification & Platform Accountability

- Enforce regulations on platforms to control the spread of deepfakes. Platforms must verify identity before content dissemination (e.g., Digital Services Act, EU guidelines).
- Develop blockchain or secure verification methods for content authenticity.

Overcome bias in deepfakes

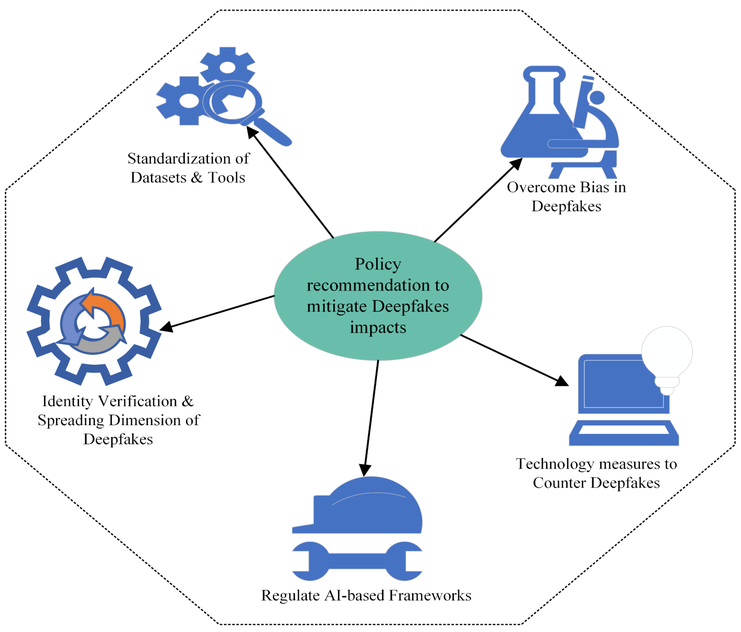
- AI-driven fakes boost microtargeting, complicating disinfo detection efforts.
- Detection lags as deepfake tools advance, dropping costs and raising threats.

Regulate AI-based Frameworks

- Establish clear roles for content creators, platforms, and regulatory bodies in preventing deepfake misuse.
- Utilize both firm and flexible regulatory measures to curb harmful content while supporting innovation.

Technological Countermeasures

- Invest in advanced AI-based detection methods.
- Protect detection tools from misuse by ensuring transparency without compromising security.



Deepfake detection and generation Landscape

Deepfake Generation Techniques

- **Face Swapping & Reenactment**: Techniques replace facial features in videos to create lifelike fakes. Challenges include realism under variable conditions (lighting, angles).
- **Synthetic Content Creation via GANs & Diffusion Models**: Advanced AI models like GANs (e.g., StyleGAN2, IC-DGAN) create high-quality synthetic images and videos that blur the line between real and fake.
- **Attribute Manipulation**: An AI-based framework can alter attributes like age, gender, and expression, creating realistic but deceptive media. Fig. 6 shows a pipeline for attribute and facial feature manipulation.

Conclusion & Future Directions

The dual nature of deepfakes demands an integrated approach combining AI innovation, ethical governance, and effective policy frameworks. Ensuring a balance between technological advancement and societal responsibility is essential in mitigating the risks of deepfakes while fostering positive use cases.

Actionable Steps

- **Research Investment:** Focus on improving detection techniques for deepfake detection in diverse real-world contexts.
- **Public Education:** Raise awareness about the societal impact of deepfakes and the importance of critical media consumption.
- **Policy Collaboration:** Work with tech companies, policymakers, and legal experts to create robust, globally coherent regulations.