Guardians of Forensic Evidence: Evaluating Analytic Systems Against AI-Generated Deepfakes

Haiying Guan, Andrew Zhang, and Jim Horan

Multimodal Information Group (MIG), Information Access Division (IAD),
Information Technology Laboratory (ITL), National Institute of Standards and Technology (NIST)

AI-generated Deepfakes: Challenges and Motivation

Growing Legal Challenges

Deepfakes pose significant challenges that could strain legal frameworks in the near future[1]

[1] Is the legal system ready for AI-generated deepfake videos? BY LAURA LOREK, AUGUST 1, 2024, American Bar Association Journal, https://www.abajournal.com/magazine/article/is-the-legal-system-ready-for-ai-generated-deepfake-videos.

Economic and Business Risks

“Across industries, businesses have lost an average of nearly $450,000 to deepfakes with 28% reporting the losses exceeded $500,000. Indeed, financial services businesses lost a little over $600,000 on average while fintech businesses lost an average of more than $630,000[2]”.

It is projected to cause global losses totaling $1 trillion by 2024[3].

Advanced Analytic Tools

are critical to preserving the integrity of forensic evidence and ensuring justice in the digital age.

AI-generated disinformation erodes public trust in media platforms and complicates forensic investigations.

Threat to Forensic Integrity and Trust in Media

Courts and law enforcement are struggling to keep pace with the rapid growth of AI-powered fraud.

Pressure on Legal and Forensic Systems

[2] ‘The Deepfake Trends 2024’. Oct. 25, 2024. https://regulaforensics.com/resources/deepfake-report-2024/.

[3] ‘AI-assisted fraud schemes could cost taxpayers $1 trillion in just 1 year, expert says’, June 20, 2023. https://www.foxnews.com/us/ai-assisted-fraud-schemes-could-cost-taxpayers-1-trillion-one-year-expert-claims.

Forensics Deepfake Evaluation program

COLLABORATION

Partner with experts from various fields, including data collection, deepfake generation, and AI analytics system development, to build the evaluation program.

ENGAGE PARTICIPANTS AND RUN THE EVALUATION

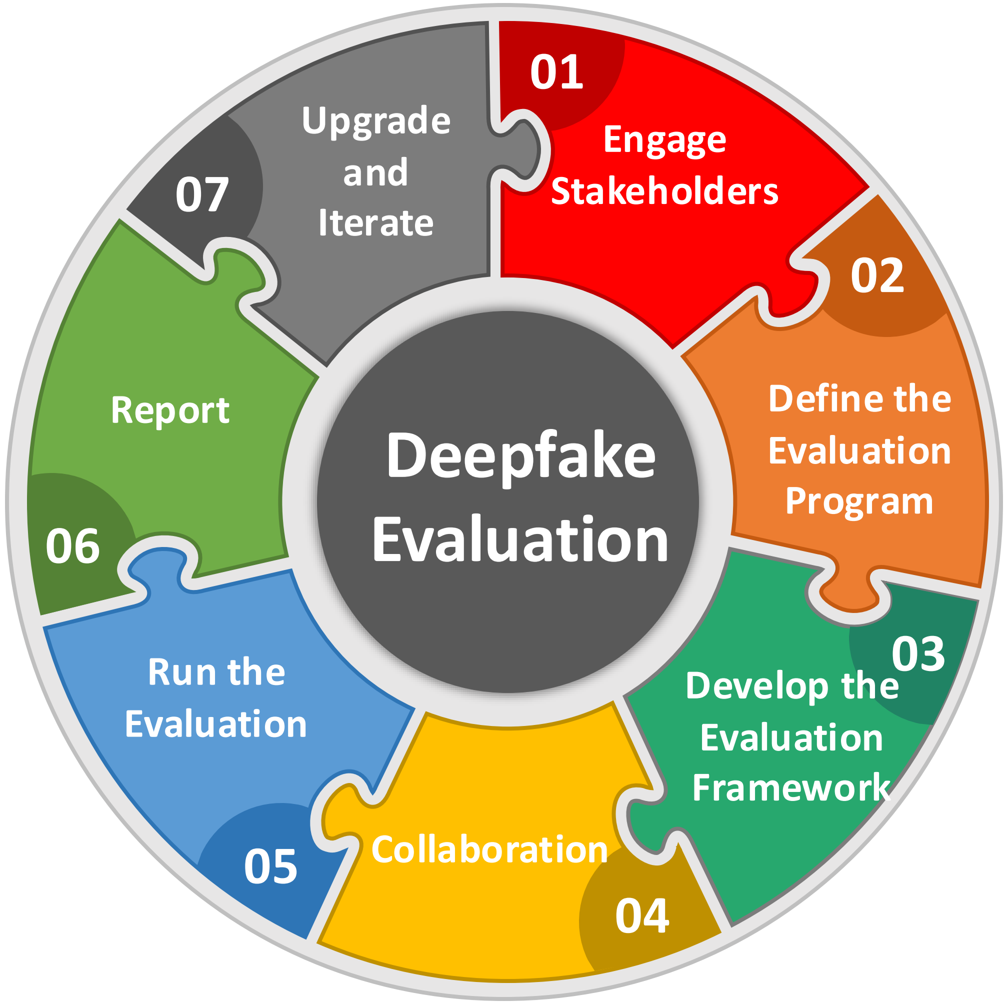
Involve relevant participants to test the AI systems. Collect feedback and performance data during the evaluation.

REPORT SYSTEM PERFORMANCE AND FINDINGS

Analyze the results and present performance metrics, strengths, weaknesses, and insights in a detailed report for stakeholders.

UPGRADE AND ITERATE

Share lessons learned, adjust evaluation task and upgrade data based on participant feedback, and continuously improve the evaluation program to integrate emerging technologies and methodologies.



ENGAGE STAKEHOLDERS

Identify and involve key stakeholders, including researchers, industry professionals, and policymakers, to define goals and expectations for the evaluation program.

DEFINE THE EVALUATION PROGRAM

Describe the program's purpose, objectives, and scope. Outline the AI systems to be evaluated and the metrics or benchmarks that will measure performance.

DEVELOP THE EVALUATION FRAMEWORK

Develop a robust evaluation framework, including methodologies, datasets, and tools. Ensure the framework accommodates both technical metrics and human-machine interaction aspects.

Program Goals

- Foster research in Deepfake and Generative AI
- Conduct recurring evaluations for state-of-the-art insights
- Collaborate with academia/industry to establish a reference baseline detection system
- Provide performance analysis for iterative system improvement
- Support the transition from lab prototypes to real-world products
- Enhance generalization of detection tools
- Deliver cross-year comparison reports

Challenges for forensics researchers

NIST

Challenges

Generalization Capability

Robustness and Resilience

Post-Processing, Laundering, anti-forensics

Data diversity and resource constraints

Rapid Evolution of Generative AI/Deepfake Technology

Evaluation Strategies

Design Evaluation to test generalization capability

Collaboration **with multiple analytic teams**

Measure the system robustness on realistic data

Release the datasets **(with IRB approval)**

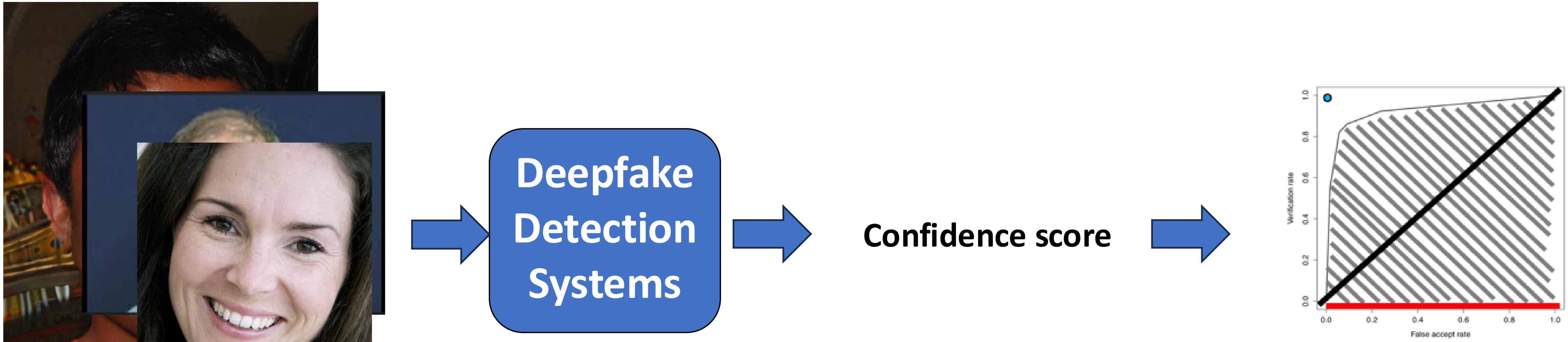
Collaboration **with multiple data teams**

Continuous Evaluation

Data generation infrastructure supports update

Forensics Deepfake Detection System Evaluation

The open evaluation program is designed to advance the development of forensic technologies for automatically detecting deepfakes and AI-generated media.



Deepfake Detection Systems

Confidence score

System Performance Reports

System Input

System

System Output

Leaderboard

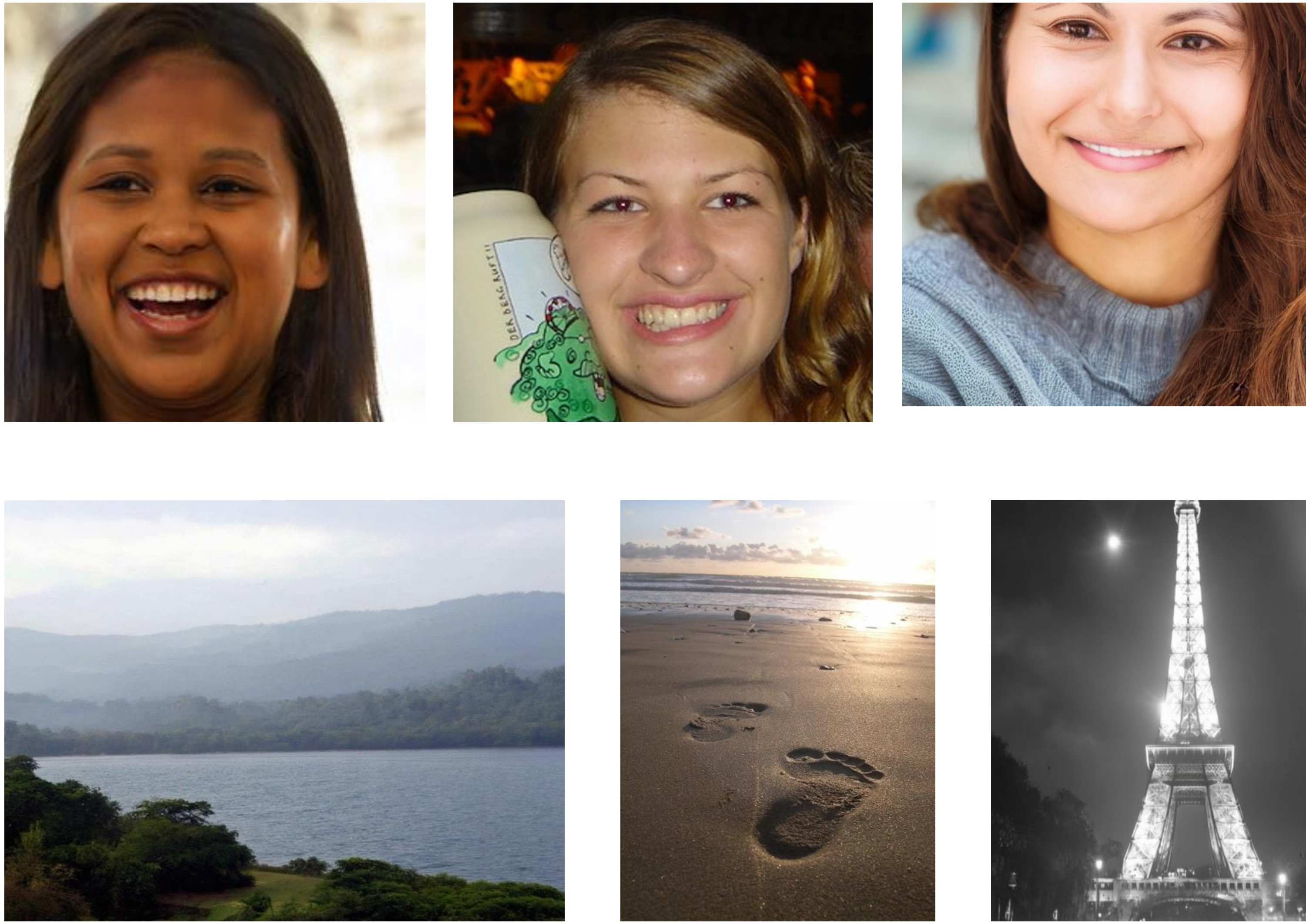
Image Deepfake Detection Task

Task: Detect deepfakes or AI-generated images.

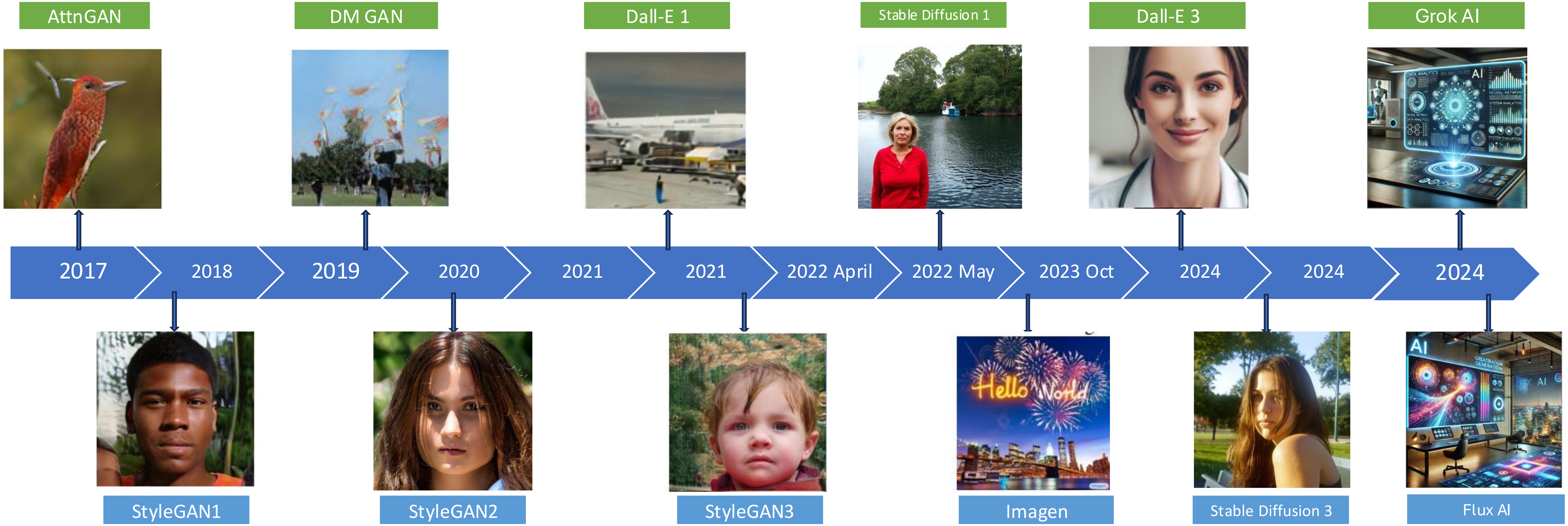
Study: Generalization Capability and Robustness

Data: StyleGANs, Stable Diffusion (SD) tools, customized tools etc.

Next Phase: Working on the evaluation dataset generation



Generative AI Tools



AttnGAN

DM GAN

Dall-E 1

Stable Diffusion 1

Dall-E 3

Grok AI

2017

2018

2019

2020

2021

2021

2022 April

2022 May

2023 Oct

2024

2024

2024

StyleGAN1

StyleGAN2

StyleGAN3

Imagen

Stable Diffusion 3

Flux AI

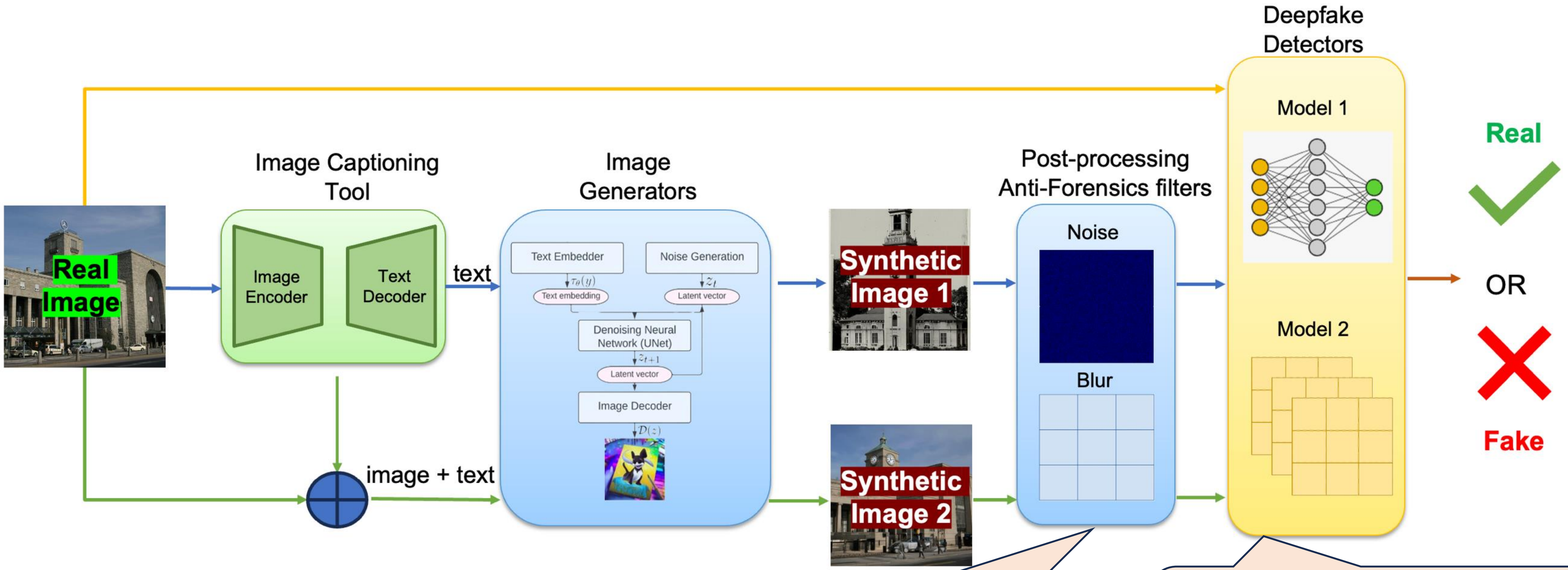
NIST Disclaimer: Certain equipment, instruments, software, or materials are identified in this presentation in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement of any product or service by NIST, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

Understanding the Gaps: A Study on Deepfake Detection Challenges

Three Challenges:

- The big gap:
  - Research algorithm accuracy: over 90%.
  - Real-world applications: social media platforms lack easy-to-use features, e.g., an add-on button.
- Generalization:
  - Systems perform well on media created by familiar generators but struggle with deepfakes generated by new or unfamiliar methods.
- Robustness:
  - In real-world applications, deepfakes often undergo post-processing.
  - The performance of algorithms on post-processed data remains uncertain.

A Study Design on the Generalization and Resilience of Deepfake Detection Systems



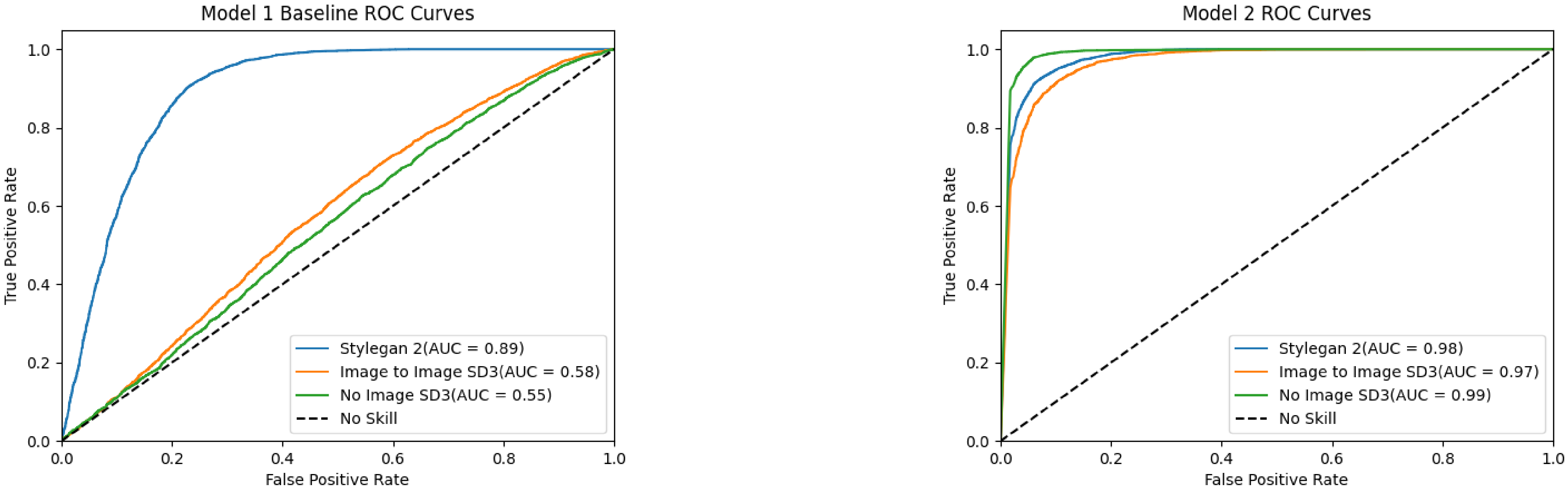
Test images are passed through various post-processing filters to evaluate resilience.

To evaluate generalization, detectors are trained on deepfakes created using older methods and tested on both older and newer deepfake generation techniques.

Generation of Image Deepfake Datasets

Generalization of Deepfake Detection Systems

- Both detection systems are trained on deepfakes generated using older methods and tested on both older and newer deepfake generation techniques.
  - The blue line represents data from known generators (older deepfake generation methods).
  - The orange and green lines represent data from unknown generators (newer deepfake generation methods).
- The results show that Detection System 1 struggles to generalize to unknown generators, while Detection System 2 performs effectively.

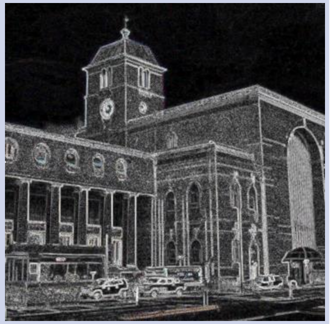
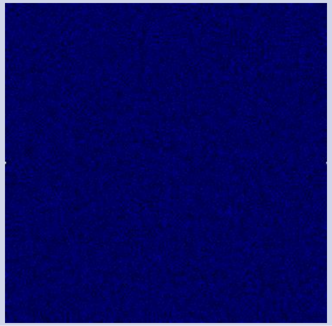
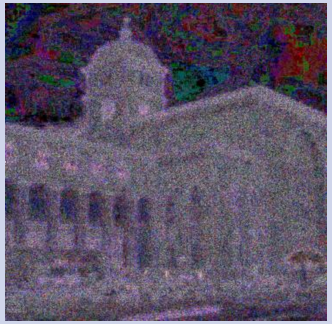
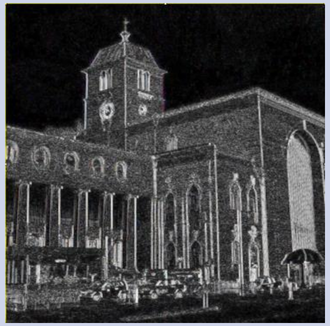
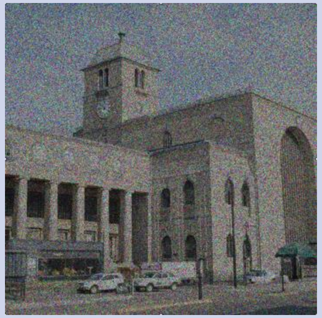
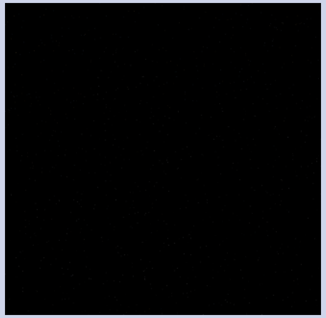
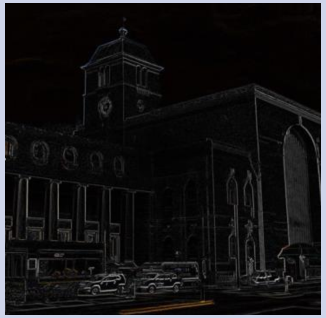


(a)

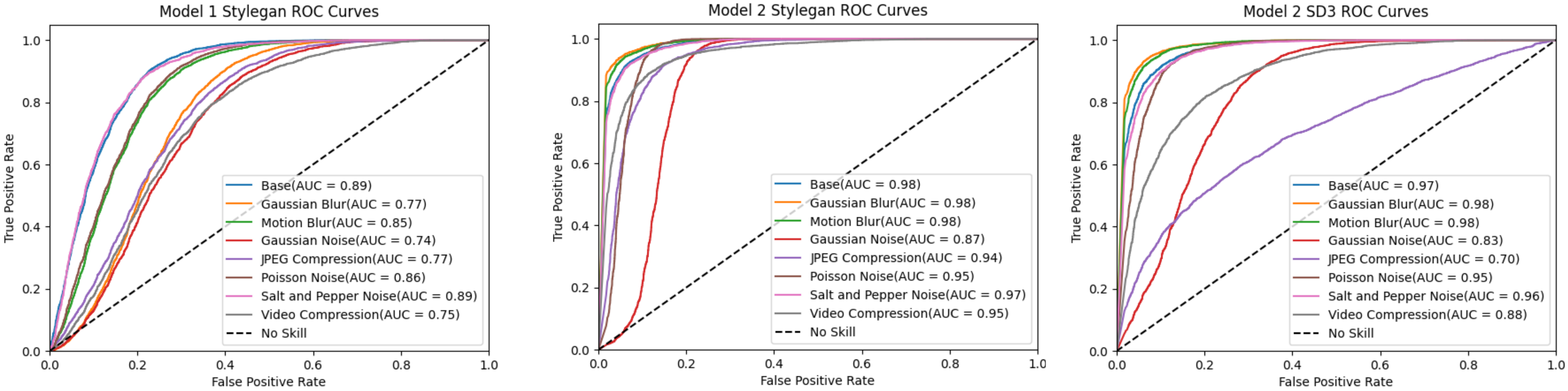
(b)

Performance of Two Image Deepfake Detection Systems on three Datasets Generated Using Different Deepfake Generation Tools

Robustness: Post-Processing Filters

	Gaussian Blur	Gaussian Noise	JPEG	Motion Blur	Poisson Noise	S&P noise	Video Compression
Original							
Processed							
Difference							

Resilience/Robustness of Deepfake Detection Systems



(a) (b) (c)

Performance of Two Image Deepfake Detection Systems on a Set of Post-Processing Datasets Using Different Deepfake Generation Tools

AFI2 Makes It Possible

AFI2 funds a portion of the work within the overall evaluation program:

- Generalization and Resilience Study
    - Supported research into generalization and resilience in deepfake detection systems, a key component for designing the evaluation program.
- Collaborations and Partnerships
    - Enabled collaborations with top academic teams
        - Professor Siwei Lyu’s team from University at Buffalo, SUNY
        - Professor Matthew Stamm’s team from Drexel University
    - Facilitated a partnership with Deep Media AI to acquire the latest synthetic and deepfake media data
- Data and Tools Enhancement
    - It is expected to provide resources to incorporate more comprehensive data and tools into future evaluations, enhancing the program's overall impact.

Challenges for the Evaluation Program

Challenges in Deepfake Detection

- Generalization Capability
- Robustness
- Rapid Advancement of Deepfake Technologies

Evaluation Program Strategies

- Collaborate with academia and industry teams to ensure representative data and data diversity.
- Work with experts in deepfake or synthetic media generation to create more realistic test images, incorporating post-processing, social media laundering, and anti-forensics filters.
- Build a flexible evaluation dataset generation infrastructure and establish partnerships with academia and industry to integrate new tools and techniques.

Acknowledgements

- External collaborators:
  - Prof. Siwei Lyu, Dr. Shan Jia, and Yan Ju at University at Buffalo
  - Prof. Matthew Stamm from Drexel University
  - Rijul Gupta team from Deep Media AI
  - Prof. Conrad Sanderson
- NIST contributors
  - Lowen DiPaula (PATHWAY)
  - Andrew Zhang (SURF)
  - Lukas Diduch
  - Ilia Ghorbanian (PREP)
  - Edmond Golden
  - John Garofolo
  - Baptiste Chocot

Thank You!

haiying.guan@nist.gov