Spotting the Unreal: Advances in Synthetic Media Detection

Luisa Verdoliva, Christos Koutlis

UNINA, CERTH

24 June 2025

vera.ai final online webinar

Overview

- Introduction to the Problem
- Challenges and Limitations
- vera.ai Research Contributions
- Tools for Synthetic Media Detection in vera.ai

Overview

- **Introduction to the Problem**
- Challenges and Limitations
- vera.ai Research Contributions
- Tools for Synthetic Media Detection in vera.ai

Fully synthetic images

First generation synthetic models - random generated image (no control) (e.g. GANs, StyleGAN)

Fully synthetic images

First generation synthetic models - random generated image (no control) (e.g. GANs, StyleGAN)

Second generation synthetic models can generate content based purely on text prompts (e.g. DALL-E, SD)

"a corgi in a field"	*"a monkey is eating a banana"*	*"A shirt with the inscription 'I love generative models' "*

“a monkey is eating a banana”

Evolution of generation control

First generation synthetic models - random generated image (no control) (e.g. GANs, StyleGAN)

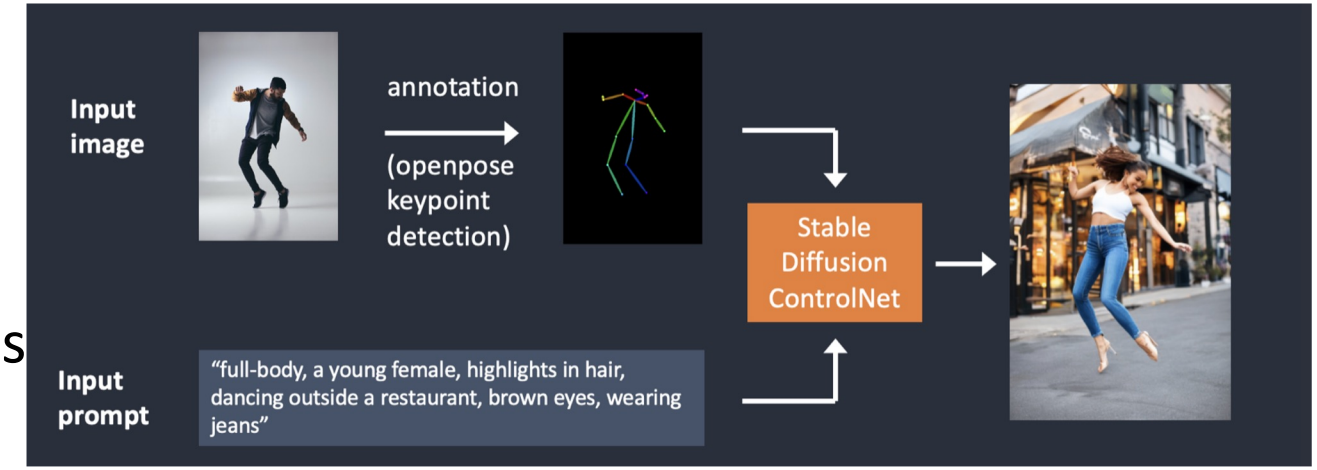
Second generation synthetic models can generate content based purely on text prompts (e.g. DALL-E, SD)

Evolution of generation control

First generation synthetic models - random generated image (no control) (e.g. GANs, StyleGAN)

Second generation synthetic models can generate content based purely on text prompts (e.g. DALL-E, SD)

Finer-grained control by providing input image and describing the modifications (e.g. ControlNet)

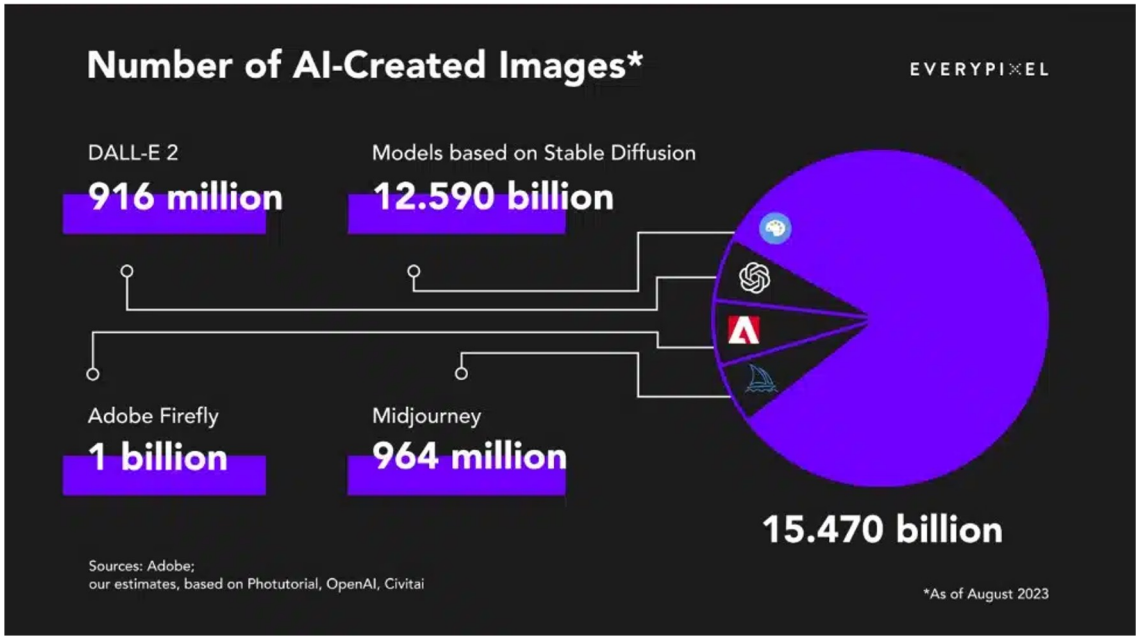


Example source: https://stable-diffusion-art.com/controlnet/

Most popular models

- Stable Diffusion → Stability AI DreamStudio, Leonardo, ...
- OpenAI DALL-E 3 → Copilot Designer
- Midjourney
- MS Designer Image Creator
- FLUX. 1
- Imagen 2 (in Gemini)
- Adobe Firefly
- NightCafe
- Starryai
- Ideogram

....



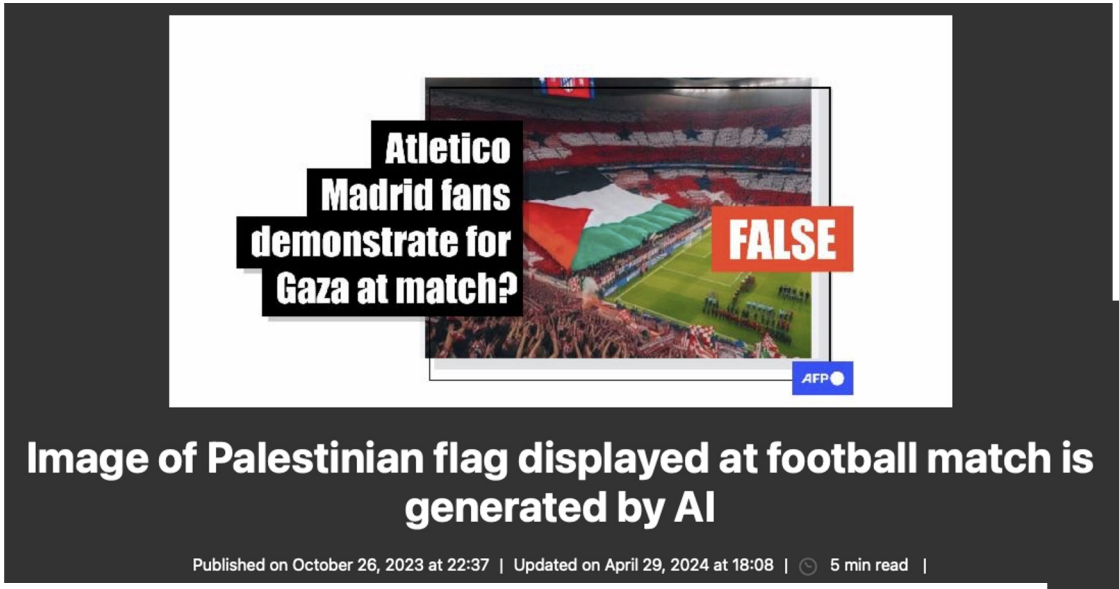
https://journal.everypixel.com/ai-image-statistics

Synthetic media is taking over the Internet…

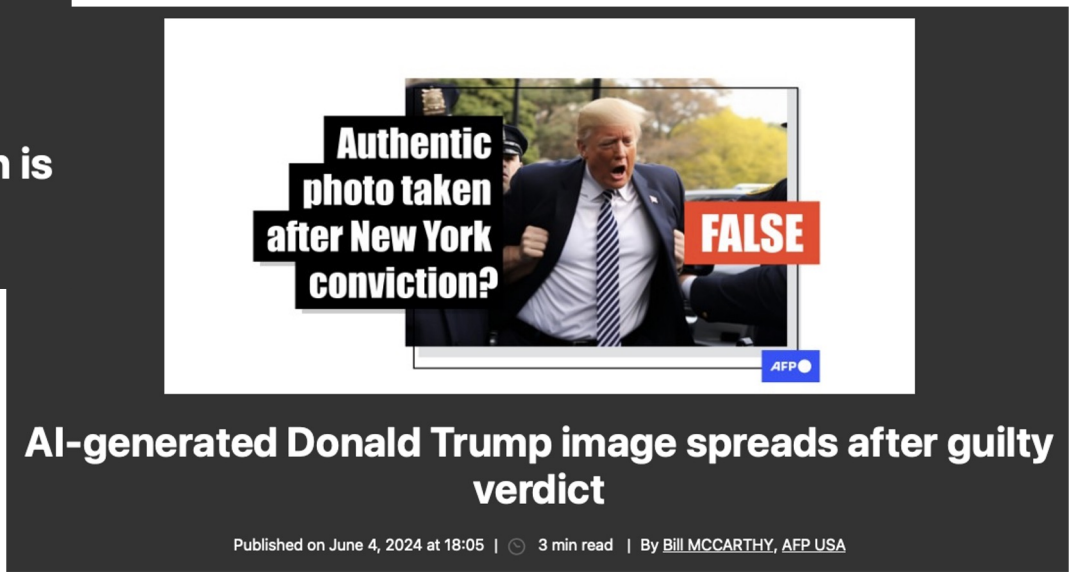
vera.ai

Real-world examples

GenAI in real news



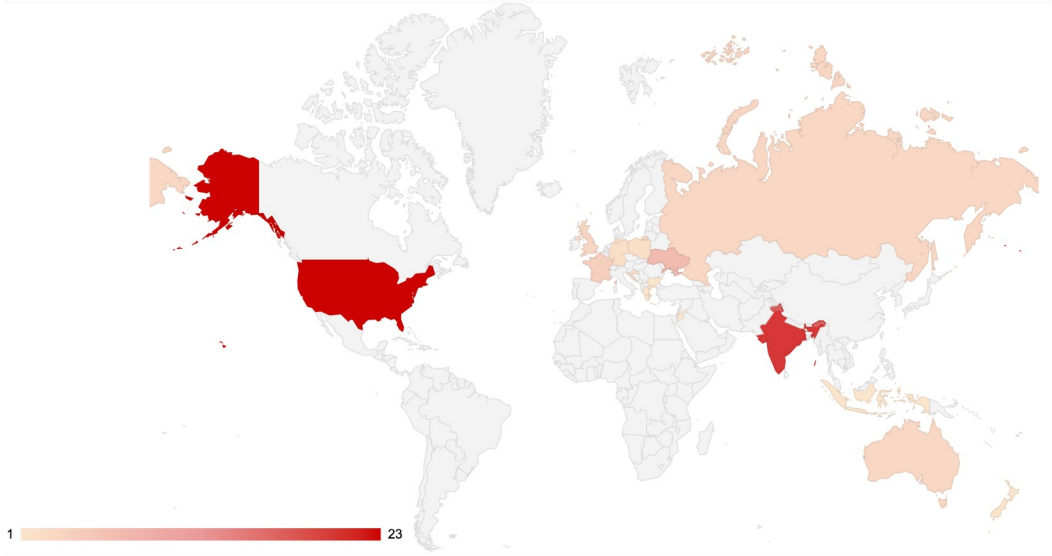
https://factcheck.afp.com/doc.afp.com.33YY7NY

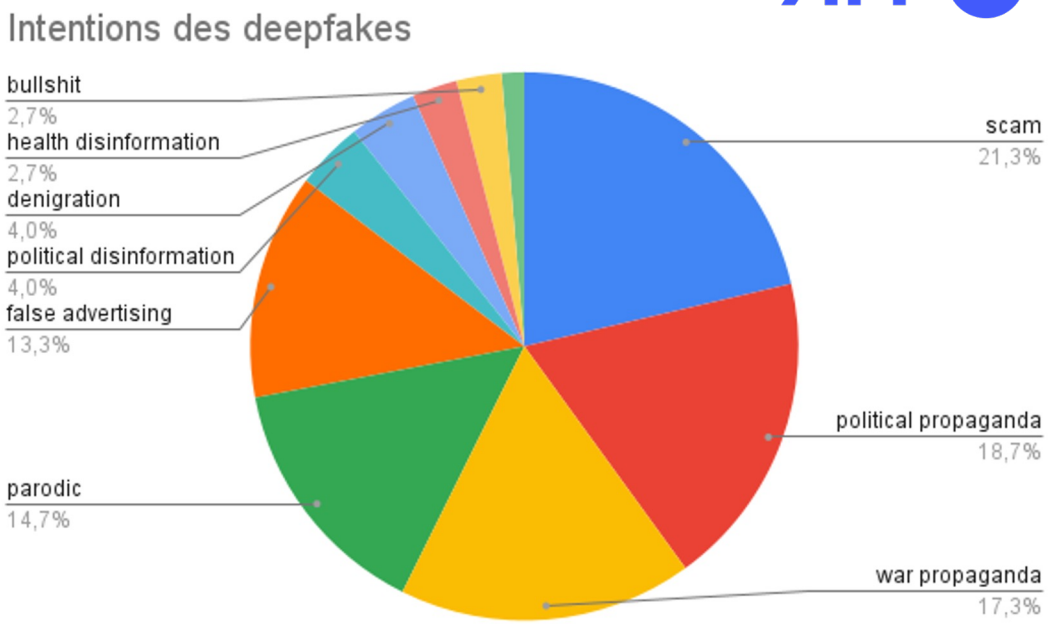


https://factcheck.afp.com/doc.afp.com.34UR4UB

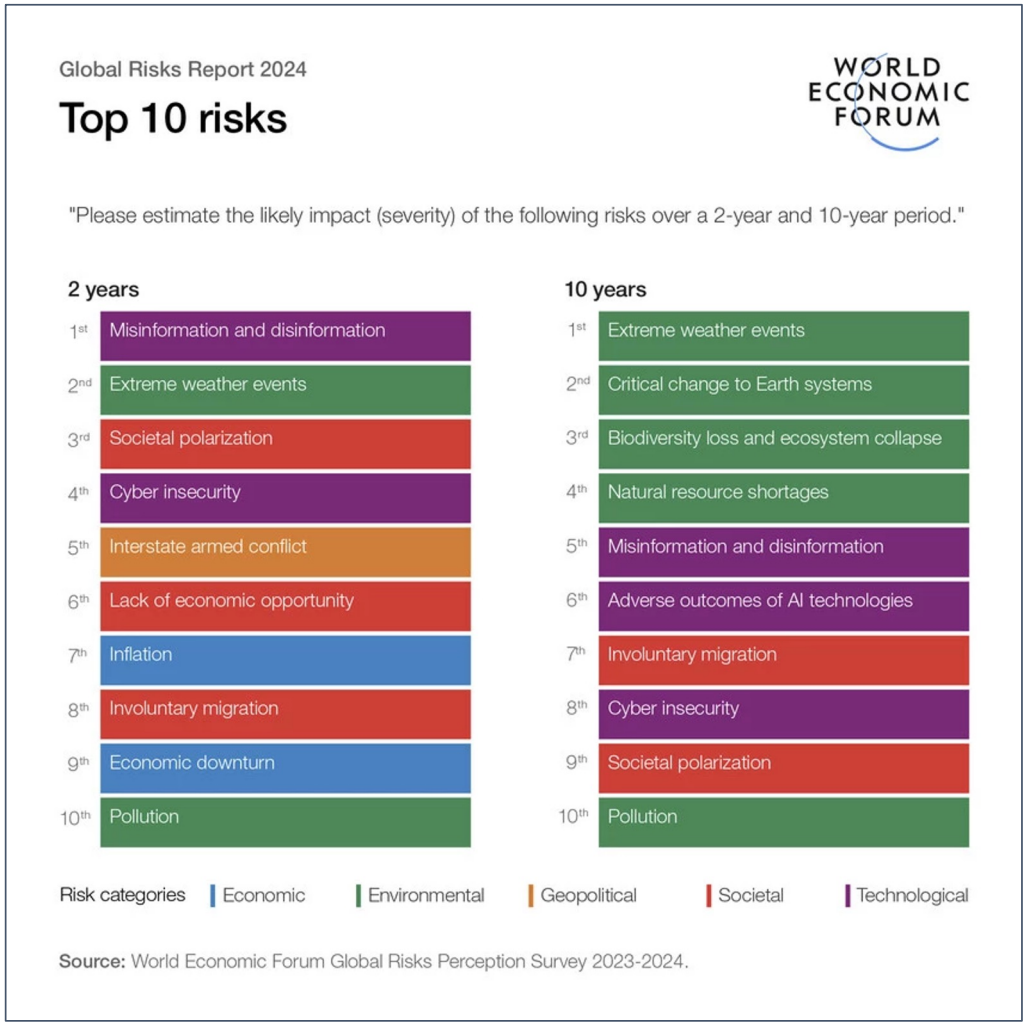
GenAI Disinformation Trends

Small-scale analysis of 75 fact-checked deepfake cases (GlobalFact 2024)





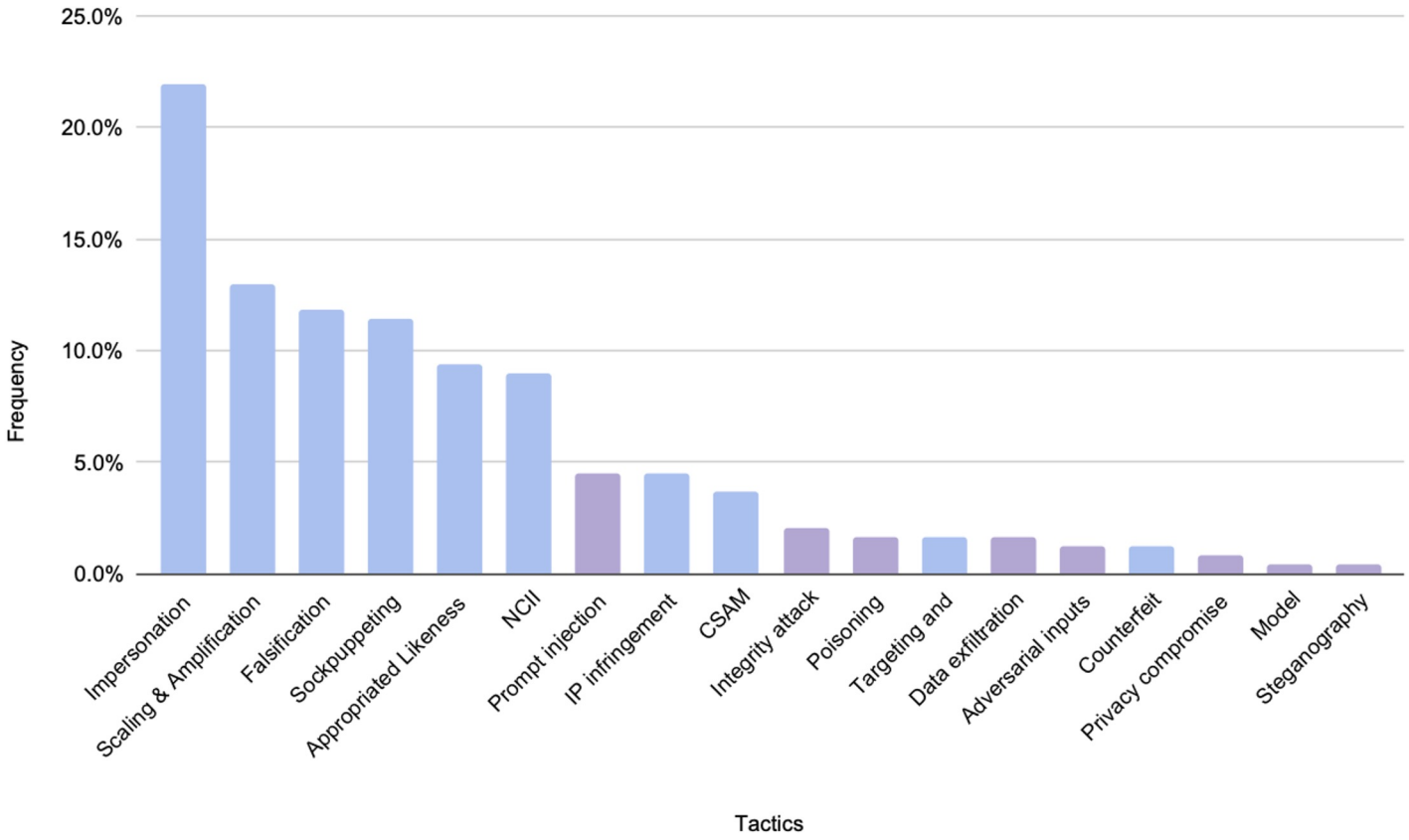
Disinformation: a pressing concern



"Disinformation in 2024 is still a pressing concern, and tops lists of short-term risks of AI, such as the World Economic Forum Global Risks report."

"Concerns over a persistent cost-of-living crisis and the intertwined risks of AI-driven misinformation and disinformation, and societal polarization dominated the risks outlook for 2024."

Insights from GenAI misuse tactics



A qualitative analysis of approximately 200 observed incidents of misuse

Tactic	Modality				Total
	Image	Text	Audio	Video	
Impersonation	4	3	28	21	56
Sockpuppeting	17	18	7	6	48
Scaling & Amplification	15	24	4	1	44
Falsification	16	12	4	2	34
NCII	11	1	1	11	24
Appropriated Likeness	12	4	2	2	20
IP Infringement	2	7	3		12
CSAM	9	1			10
Targeting/ Personalisation		5	2		7
Counterfeit		3			3
Total	86	78	51	43	258

Marchal et al., 2024, https://arxiv.org/pdf/2406.13843

…often with grave impact

Fiery explosion outside the Pentagon on May 22, 2023

https://factcheck.afp.com/doc.afp.com.33FV4BU

Incident: Social media users are claiming an image shows a fiery explosion outside the Pentagon on May 22, 2023. The Defense Department confirmed to AFP that there was no such attack, and the picture of the supposed blast appears to have been generated using AI technology.

Impact: The earliest tweet* AFP found sharing the image came from "CBKNEWS," a QAnon-promoting account that has previously shared other disinformation, though the original source of the image was not immediately known. The spread of the alleged photo appeared to cause a brief dip on Wall Street on May 22, with the S&P 500 stumbling by 0.29 percent before recovering. It also prompted live television coverage from an Indian news organization, according to journalists who shared recordings of the segment.

Overview

- Introduction to the Problem
- **Challenges and Limitations**
- vera.ai Research Contributions
- Tools for Synthetic Media Detection in vera.ai

Quality of synthetic images over time

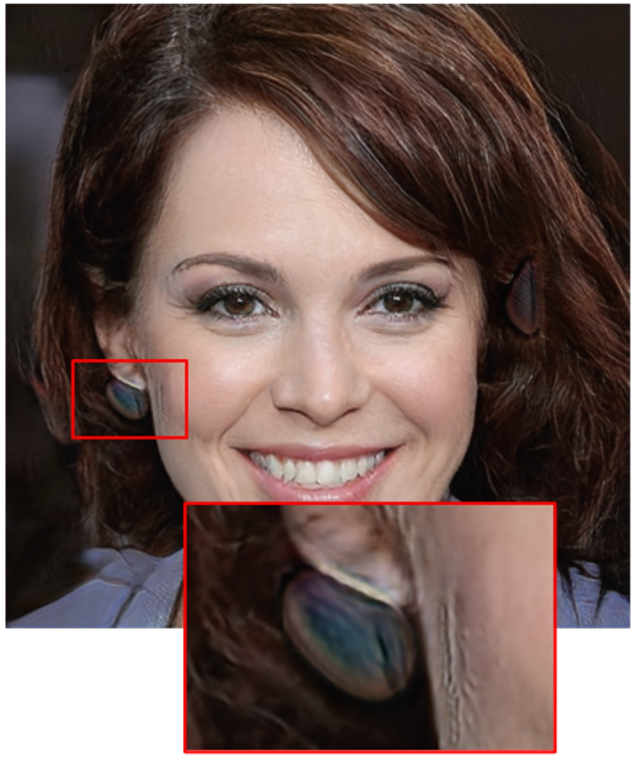
Evolution of image quality

Evolution of image quality in Midjourney

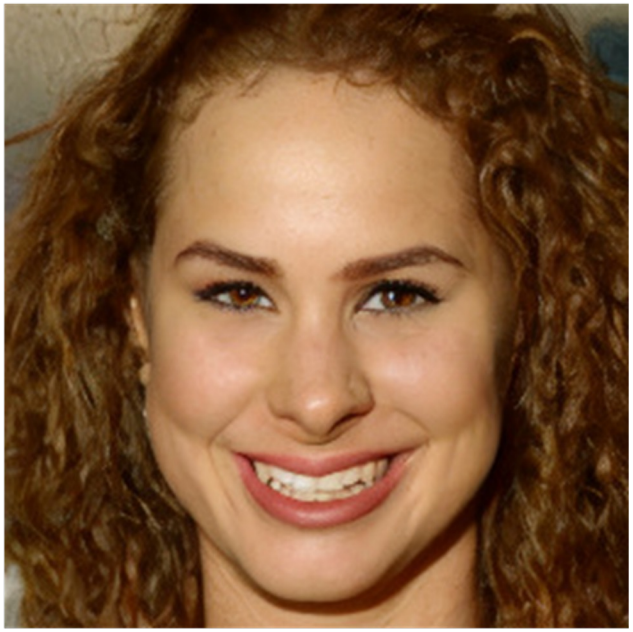
Visible anomalies

- Visible anomalies, present in images generated a few years ago, are no longer present in the latest generation techniques

2018



2022



Visible anomalies

- Visible anomalies, present in images generated a few years ago, are no longer present in the latest generation techniques

Generated Image

Shadow Errors

Vanishing Point Errors

Sarkar et al. "Shadows Don't Lie and Lines Can't Bend! Generative Models don't know Projective Geometry... for now." CVPR 2024

Highly realistic examples

vera.ai







Are forensic detectors effective?

vera.ai

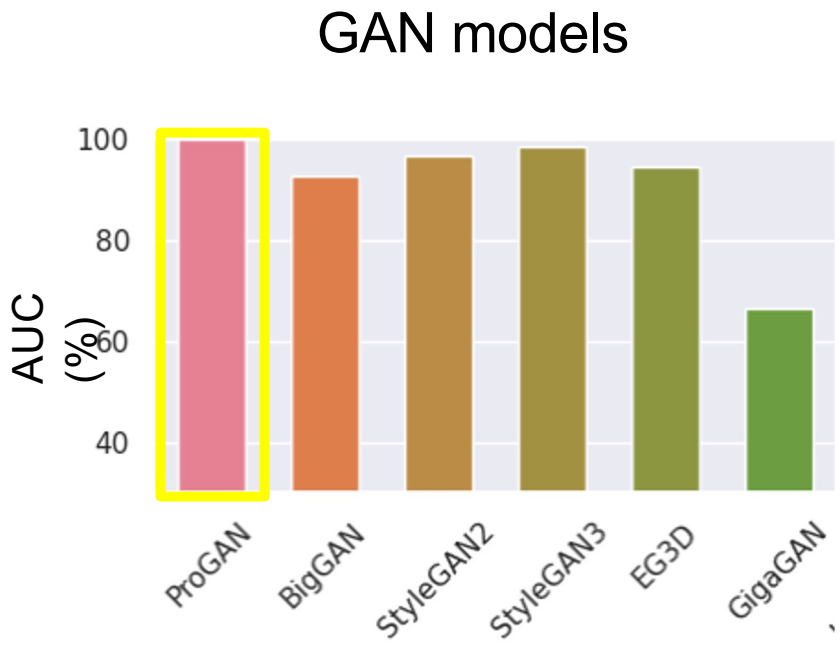






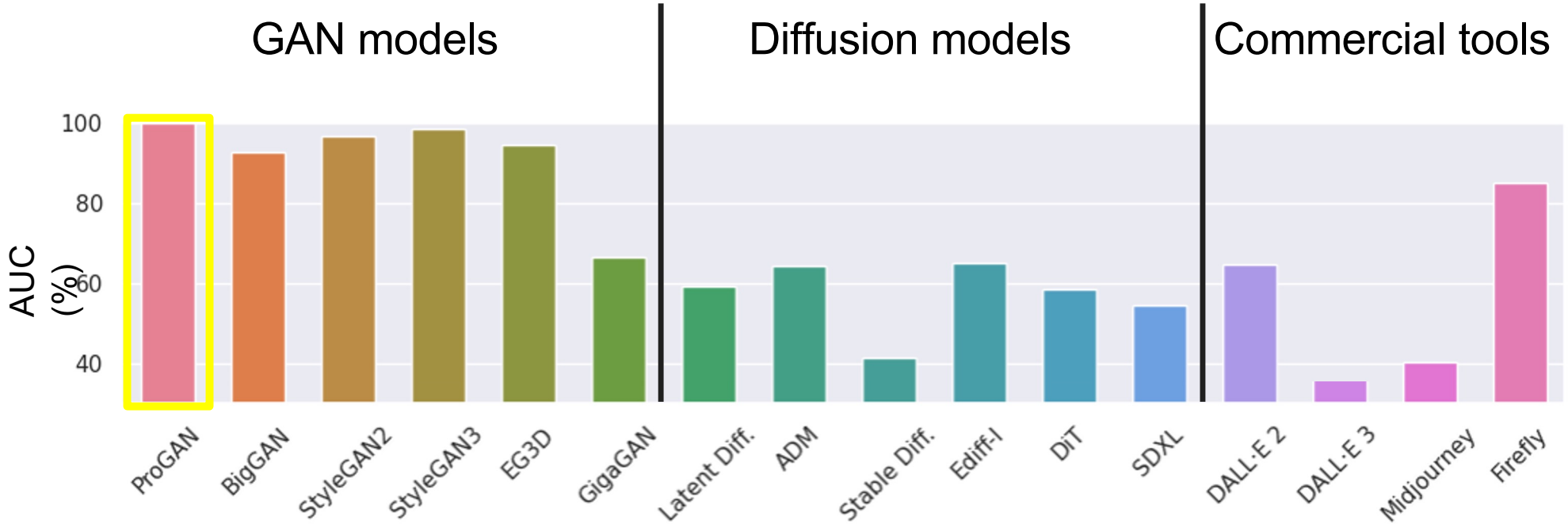
Forensic detectors depend on the training data

- Training: ProGAN



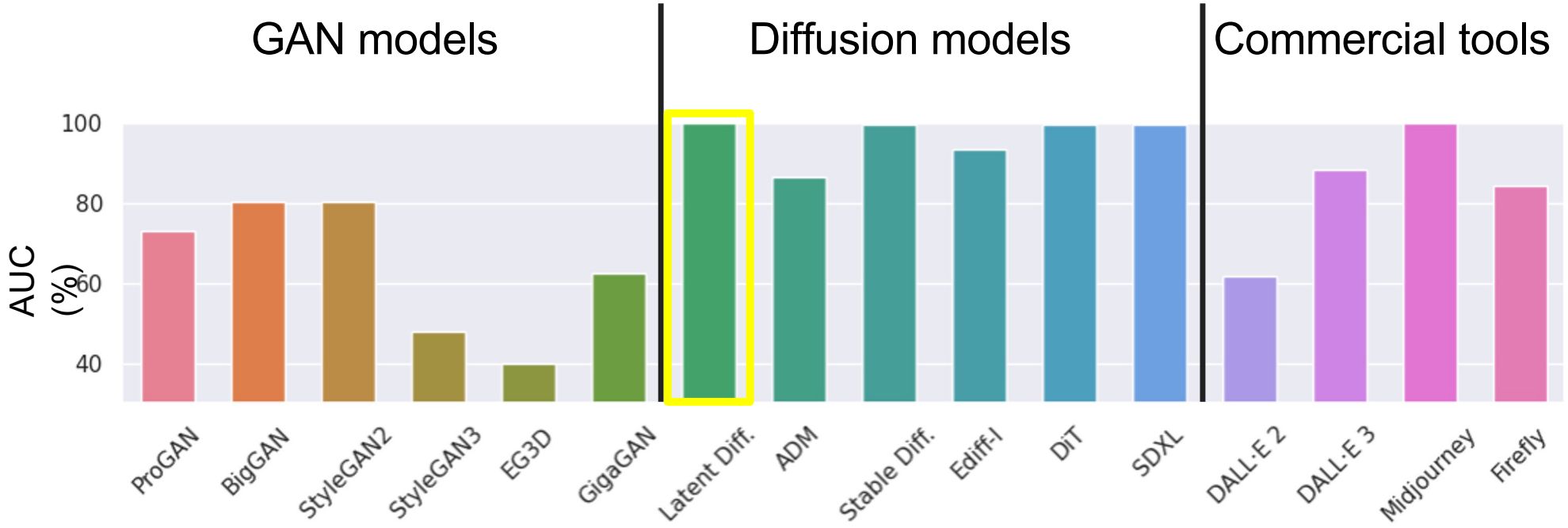
Forensic detectors depend on the training data

- Training: ProGAN



Forensic detectors depend on the training data

- Training: Latent Diffusion Model



Bias in training datasets

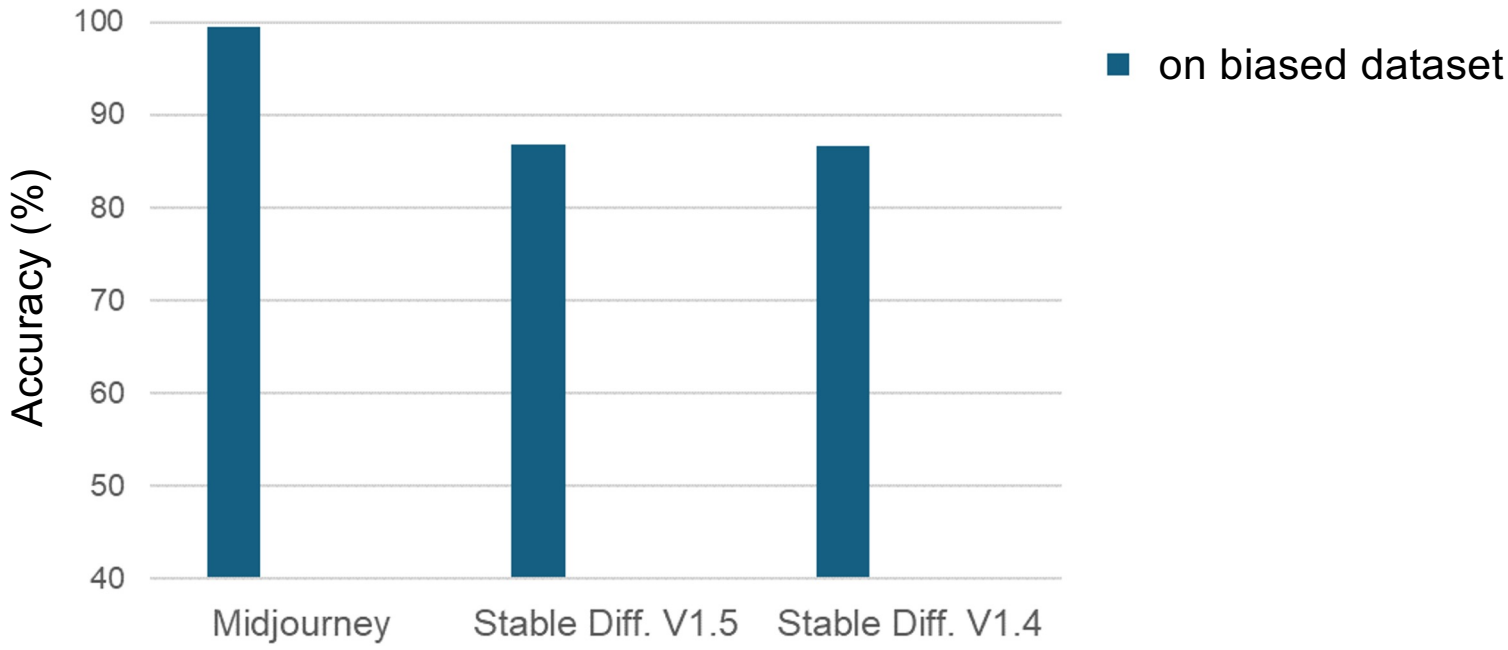
- Real images are often JPEG compressed while synthetic images are not (PNG)

Source	Real	Fake
LSUN, ImageNet, CelebA, Coco	originally JPEG	PNG
LSUN-bed	JPEG	PNG
LSUN, CelebA-HQ, ImageNet	JPEG	PNG
LAION	JPEG	PNG
ImageNet, LAION	JPEG	PNG
ImageNet	JPEG	PNG

Grommelt, et al. "Fake or JPEG? Revealing Common Biases in Generated Image Detection Datasets." *arXiv preprint arXiv:2403.17608*, 2024

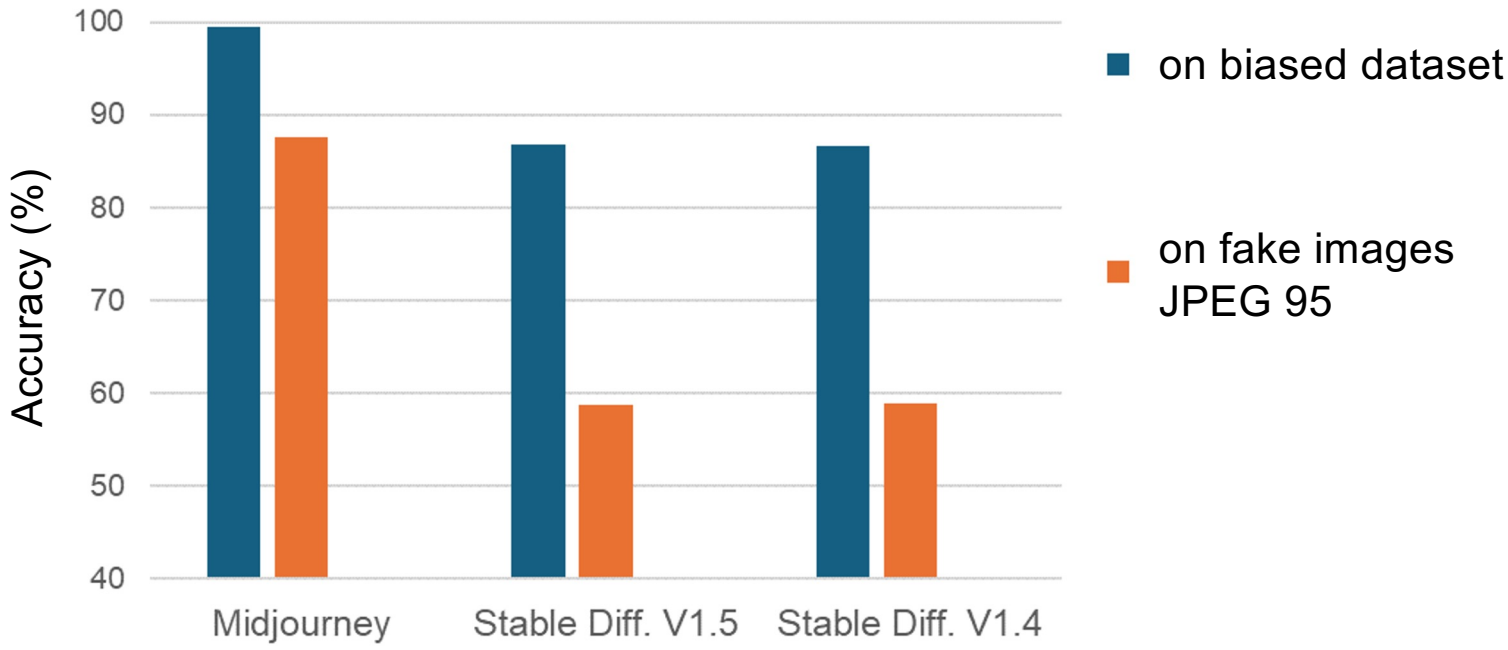
Bias in training datasets

- Results on GenImage dataset using Midjourney images



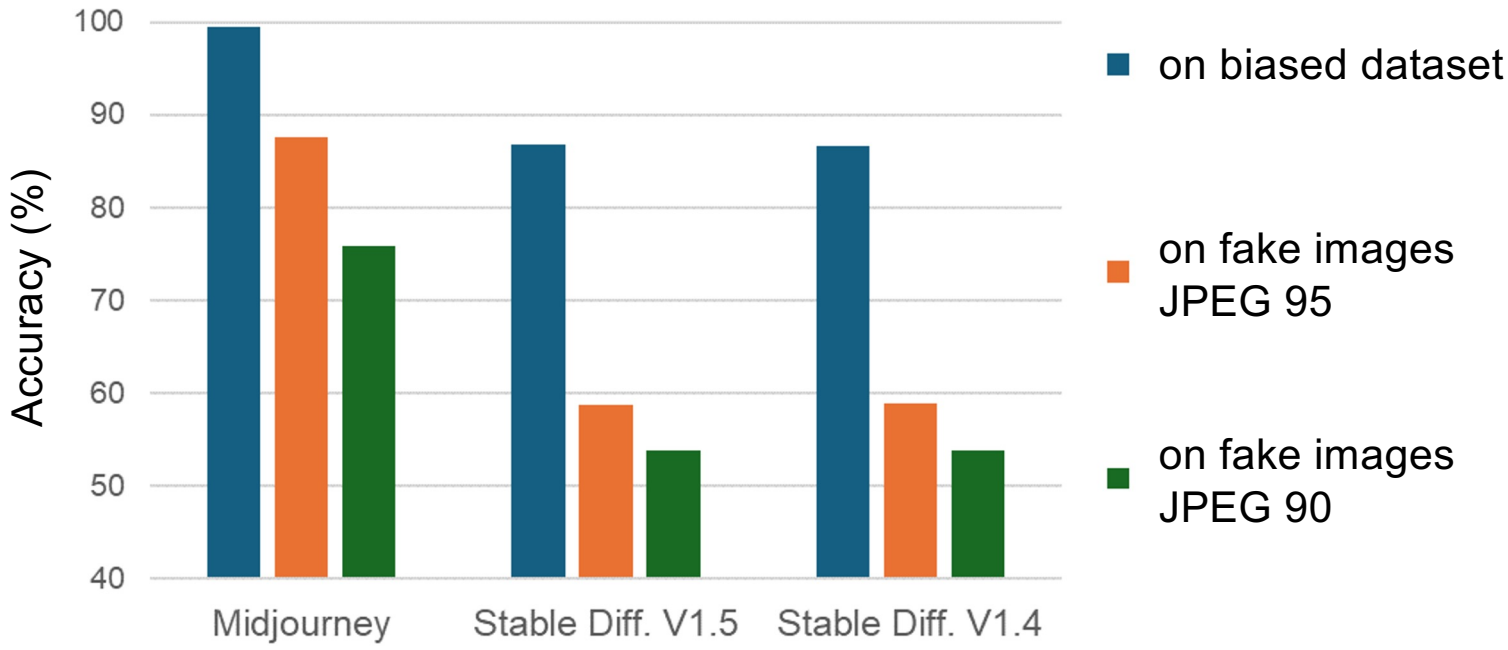
Bias in training datasets

- Results on GenImage dataset using Midjourney images



Bias in training datasets

- Results on GenImage dataset using Midjourney images



Detection challenges - Derivative Images

- Detection methods are developed to detect “base” images, i.e. images that look like actual photos.
- Image content often circulates online in the form of “derivative” images (inclusion in memes and screenshots, addition of synthetic text, image in a photo, etc.)
- In many cases SID algorithms detect the later post-processing operations, instead of the actual signal of the image, plausibly increasing performance and causing several false-positives.
- In our recent study, **considering only a subset of “base” synthetic images collected in the wild, decreases the average performance of 12 popular SID approaches by 6% in terms of AUC.**

D. Karageorgiou, Q, Bammey, V. Porcellini, B. Goupil, D. Teyssou, S Papadopoulos, “Evolution of Detection Performance throughout the Online Lifespan of Synthetic Images.”, European Conference on Computer Vision (ECCV) Workshops, 2024

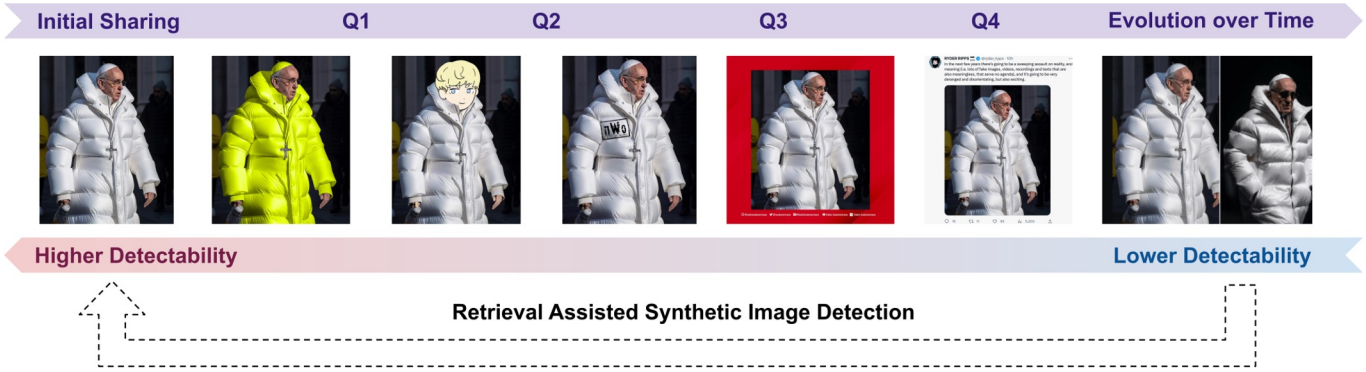
Evolution of Detection Performance throughout the Online Lifespan of Synthetic Images

Performance of popular SID algorithms on synthetic images collected in the wild.

Some perform even worse than random guessing!

Algorithm	B. Accuracy
GramNet	43.2
UnivFD	45.8
PatchCraft	47.9
Fusing	49.3
CNNDetect	49.4
FreqDetect	49.5
Dire	51.5
DIMD	53.0
NPR	59.2
Rine	61.8
LGrad	68.1
DeFake	72.8

- While state-of-the-art methods exhibit strong performance on lab-generated data, **they fail to discriminate between synthetic and real image cases collected in the wild, without further preconditions.**
- An image is continuously post-processed and reshared after its initial online appearance, leading to an average **3.2% drop in AUC between Q1 and Q4, even when considering only base images.**
- **Retrieval Assisted Synthetic Image Detection** exploits the early copies of an image submitted to a detection system, to facilitate the detection of the heavy post-processed late ones. **Increases AUC performance by 7.8% on average**.



D. Karageorgiou, Q, Bammey, V. Porcellini, B. Goupil, D. Teyssou, S Papadopoulos, “Evolution of Detection Performance throughout the Online Lifespan of Synthetic Images.”, European Conference on Computer Vision (ECCV) Workshops, 2024

Overview

- Introduction to the Problem
- Challenges and Limitations
- **vera.ai Research Contributions**
- Tools for Synthetic Media Detection in vera.ai

AI-Generated Image Detection (AIGID) Research in the vera.ai Project

Several works on AIGID have been conducted in the framework of the vera.ai project, addressing:

- Benchmarking and artifact analysis
- Leveraging foundation models
- Spectral analysis
- Generalization and bias

Benchmarking and artifact analysis

SIDBench: A Python Framework for Reliably Assessing Synthetic Image Detection Methods

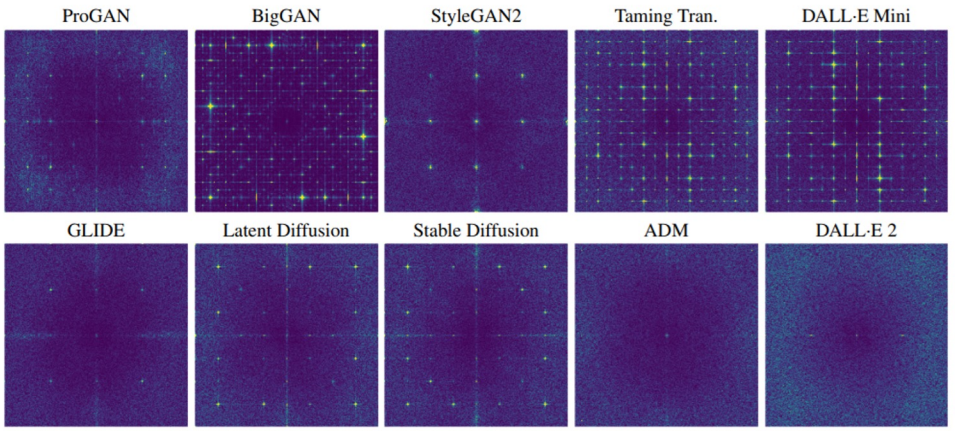
- Evaluates the effectiveness and robustness of synthetic image detection models
- Integrates 11 state-of-the-art AIGID models identified through systematic literature review
- Current AIGID methods excel at identifying images from specific generative models
- A significant hurdle is the poor generalization to new or unseen generative models
- The field faces an ongoing "arms race" to develop truly universal detectors that can identify images from any generator

Schinas, M., & Papadopoulos, S. (2024, June). SIDBench: A Python framework for reliably assessing synthetic image detection methods. In Proce
of the 3rd ACM International Workshop on Multimedia AI against Disinformation (pp. 55-64).

On the detection of synthetic images generated by Diffusion Models

- This work aims to understand the fundamental differences between real and DM-generated images
- Do DMs leave hidden artifacts similar to those found in GAN images?
- We assessed performance under real-world conditions (compr./resizing)

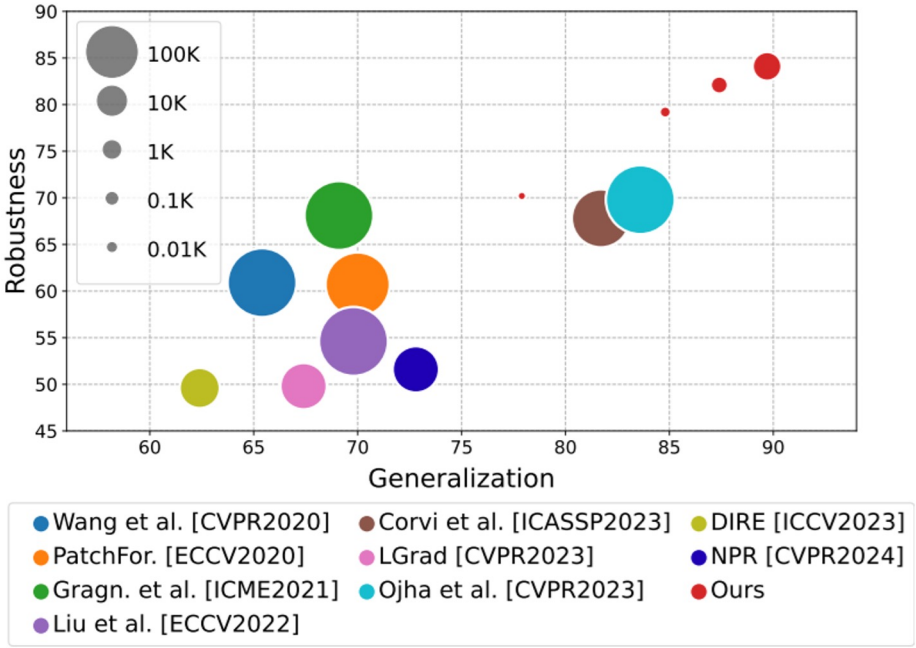
	Trained on ProGAN							
	Uncompressed				Resized and Compressed			
Acc./AUC%	Spec [87]	PatchFor. [11]	Wang2020 [79]	Grag2021 [30]	Spec [87]	PatchFor. [11]	Wang2020 [79]	Grag2021 [30]
ProGAN	83.5/ 99.2	64.9/ 97.6	99.9/100	99.9/100	49.7/ 48.5	50.4/ 65.3	99.7/100	99.9/100
StyleGAN2	65.3/ 72.0	50.2/ 88.3	74.0/ 97.3	98.1/ 99.9	51.8/ 50.5	50.8/ 73.6	54.8/ 85.0	63.3/ 94.8
StyleGAN3	33.8/ 4.4	50.0/ 91.8	58.3/ 95.1	91.2/ 99.5	52.9/ 51.9	50.2/ 76.7	54.3/ 86.4	58.3/ 94.4
BigGAN	73.3/ 80.5	52.5/ 85.7	66.3/ 94.4	95.6/ 99.1	52.1/ 52.2	50.5/ 58.8	55.4/ 85.9	79.0/ 99.1
EG3D	80.3/ 89.6	50.0/ 78.4	59.2/ 96.7	99.4/100	58.9/ 60.6	49.8/ 81.9	52.1/ 85.1	56.8/ 96.6
Taming Tran.	79.6/ 86.6	50.5/ 69.4	51.2/ 66.5	73.5/ 96.6	49.0/ 49.1	50.0/ 64.1	50.5/ 71.0	56.2/ 94.3
DALL·E Mini	80.1/ 88.1	51.5/ 82.2	51.7/ 60.6	70.4/ 95.6	59.1/ 61.9	50.1/ 68.7	51.1/ 66.2	62.3/ 95.4
DALL·E 2	82.0/ 93.0	50.0/ 51.1	50.3/ 85.8	51.9/ 94.9	61.8/ 64.5	49.8/ 58.3	49.9/ 46.1	50.0/ 65.4
GLIDE	73.4/ 81.9	50.3/ 96.6	51.1/ 62.6	58.6/ 86.4	53.1/ 52.5	51.0/ 71.5	50.3/ 65.9	51.8/ 90.0
Latent Diff.	72.1/ 78.5	51.8/ 84.3	51.0/ 62.5	58.2/ 91.5	47.9/ 46.3	50.6/ 65.2	50.7/ 69.1	52.4/ 89.4
Stable Diff.	66.8/ 74.7	50.8/ 85.0	50.9/ 65.9	62.1/ 92.9	46.5/ 44.5	51.1/ 77.2	50.7/ 72.9	58.1/ 93.7
ADM	55.1/ 53.3	50.4/ 87.1	50.6/ 56.3	51.2/ 57.4	49.1/ 49.1	51.0/ 69.1	50.3/ 68.1	50.6/ 77.2
AVG	70.5/ 75.2	51.9/ 83.2	59.5/ 78.6	75.8/ 92.8	52.7/ 52.7	50.4/ 69.2	55.8/ 75.1	61.5/ 90.8



Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., & Verdoliva, L. (2023, June). On the detection of synthetic images generated by diffusion models. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

Leveraging foundation models

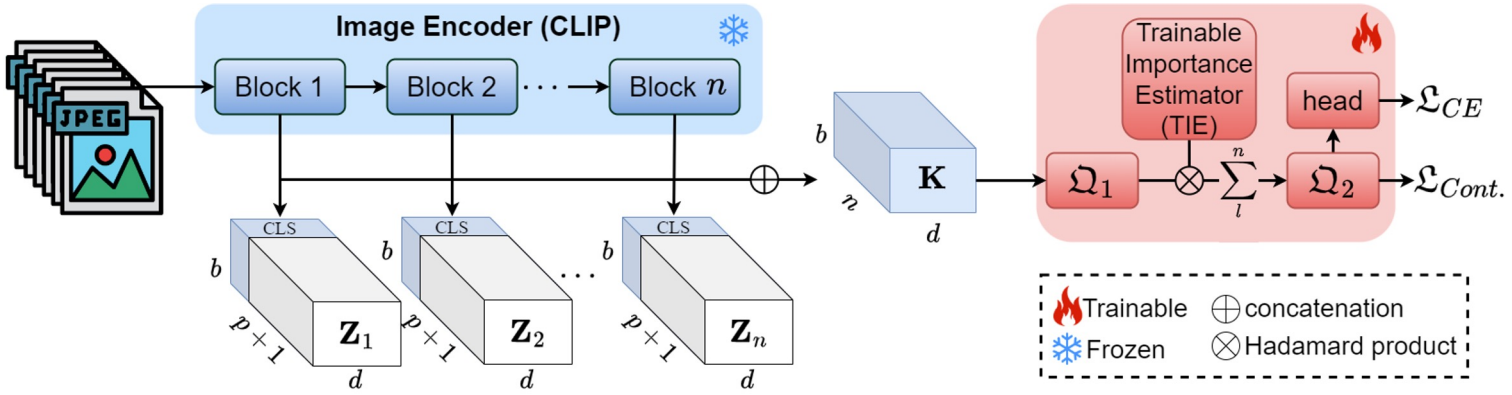
A CLIP-based model trained on a small dataset of Latent Diffusion generated images



- A lightweight, CLIP-based detector that learns robust representations from a small dataset (e.g., Latent Diffusion)
- Uses paired real/fake images (same caption) to train a linear SVM classifier
- Outperforms SoTA methods on unseen GANs, DMs, and commercial tools, even with minimal training data
- Maintains high performance on post-processed (corrupted) images

Cozzolino, D., Poggi, G., Corvi, R., Nießner, M., & Verdoliva, L. (2024). Raising the Bar of AI-generated Image Detection with CLIP. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4356-4366).

Representations from Intermediate Encoder-blocks



- Leveraging low-level visual information from intermediate Transformer blocks
- Learning a forgery-aware vector space on top of CLIP's image representations
- Training on ProGAN data & evaluation on GAN, Diffusion outperforms SotA
- Performance:
  96% (DALL-E 2), 99% (SDXL), 96% (Glide), 93% (Firefly), 97% (Midjourney)

Koutlis, C., & Papadopoulos, S. (2024, September). Leveraging representations from intermediate encoder-blocks for synthetic image detection. In European Conference on Computer Vision (pp. 394-411). Cham: Springer Nature Switzerland.

Spectral analysis

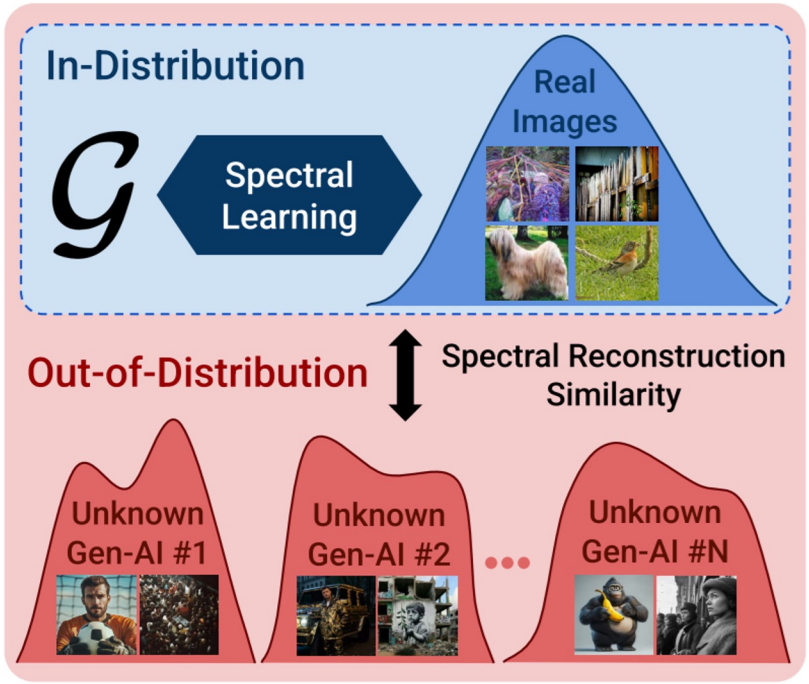
Towards spectral detection of synthetic images

- **Synthbuster** applies a high-pass filter to highlight periodic artifacts and identifies magnitudes at hard-coded peak locations (e.g., 2, 4, 8 periodicities) in the Fourier transform
- **MaskSim** replaces hard-coding with a flexible mask to identify relevant frequency components and compares the masked spectrum to learned real/fake references
- Both spectral methods (Synthbuster, MaskSim) outperform other SotA detectors on DM images, achieving high AUCs on distorted data (e.g., MaskSim at 95.5% AUC for Q=70 JPEG)

Bammey, Q. (2023). Synthbuster: Towards detection of diffusion model generated images. IEEE Open Journal of Signal Processing, 5, 1-9.

Li, Y., Bammey, Q., Gardella, M., Nikoukhah, T., Morel, J. M., Colom, M., & Von Gioi, R. G. (2024). MaskSim: Detection of synthetic images by masked spectrum similarity analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3855-3865).

Any-Resolution AI-Generated Image Detection by Spectral Learning

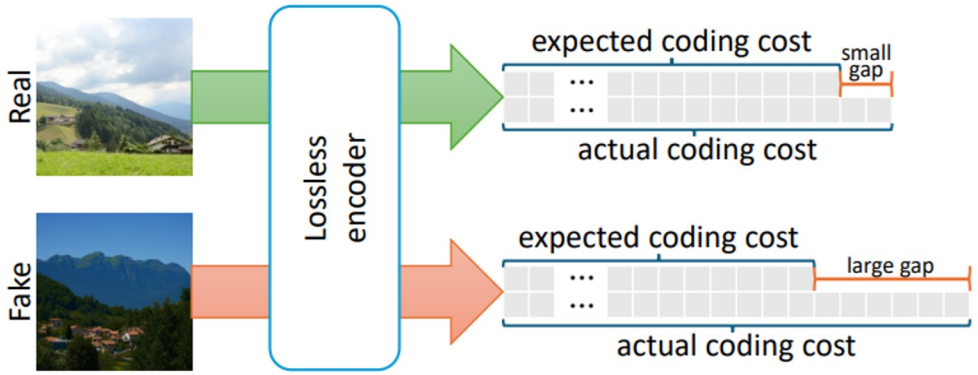


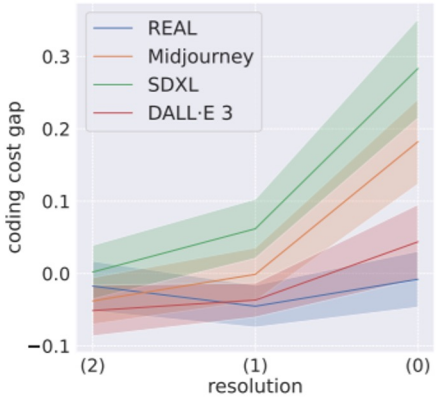
- SPAI learns the natural spectral distribution of real images in a self-supervised way
- It then identifies AI-generated images as "out-of-distribution" samples from this learned "real" model
- Pre-trained on only real images (ImageNet) and a small LDM dataset, SPAI significantly outperforms other methods on diverse, unseen generative models, including commercial tools

Karageorgiou, D., Papadopoulos, S., Kompatsiaris, I., & Gavves, E. (2025). Any-resolution ai-generated image detection by spectral learning. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 18706-18717).

Generalization and bias

Zero-Shot Detection of AI-Generated Images

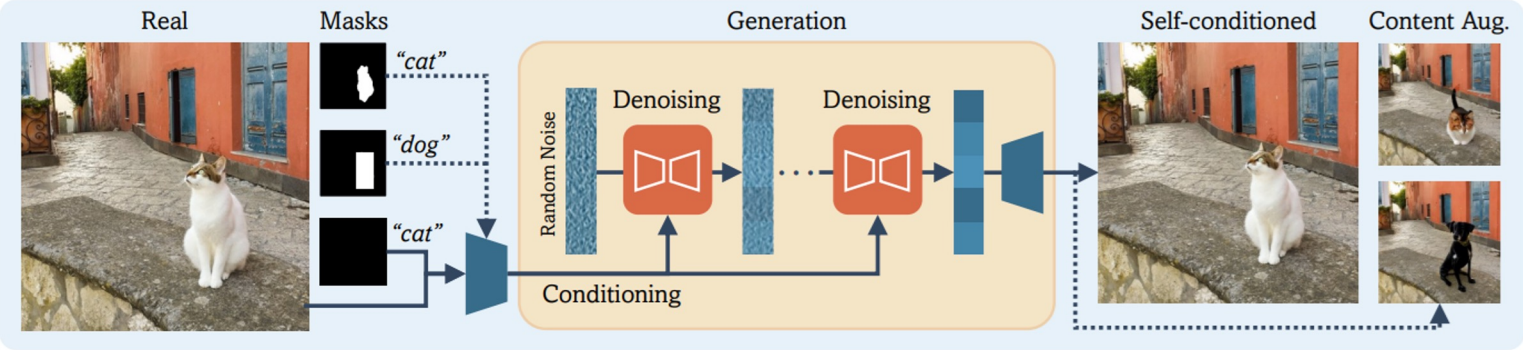




- A zero-shot detector learns a statistical model of real images
- Uses a lossless image coder (SReC) to predict pixel values at multiple resolutions
- AI-generated images are identified as outliers to the real-image model
- ZED performs on par with or better than supervised SotA, despite requiring no training on synthetic images

Cozzolino, D., Poggi, G., Nießner, M., & Verdoliva, L. (2024, September). Zero-shot detection of ai-generated images. In European Conference on Computer Vision (pp. 54-72). Cham: Springer Nature Switzerland.

A Bias-Free Training Paradigm for More General AI-generated Image Detection



- Created synthetic training images starting from real images to ensure perfect semantic alignment
- B-Free achieves high performance across current and emerging generative models (Midjourney, SDXL, DALL·E 2/3, Firefly, FLUX, SD 3.5)
- Maintains strong performance on images shared across social media (Facebook, Reddit, Twitter) and sustains accuracy over time even as images are reposted multiple times, where other detectors degrade

Guillaro, F., Zingarini, G., Usman, B., Sud, A., Cozzolino, D., & Verdoliva, L. (2025). A bias-free training paradigm for more general ai-generated image detection. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 18685-18694).

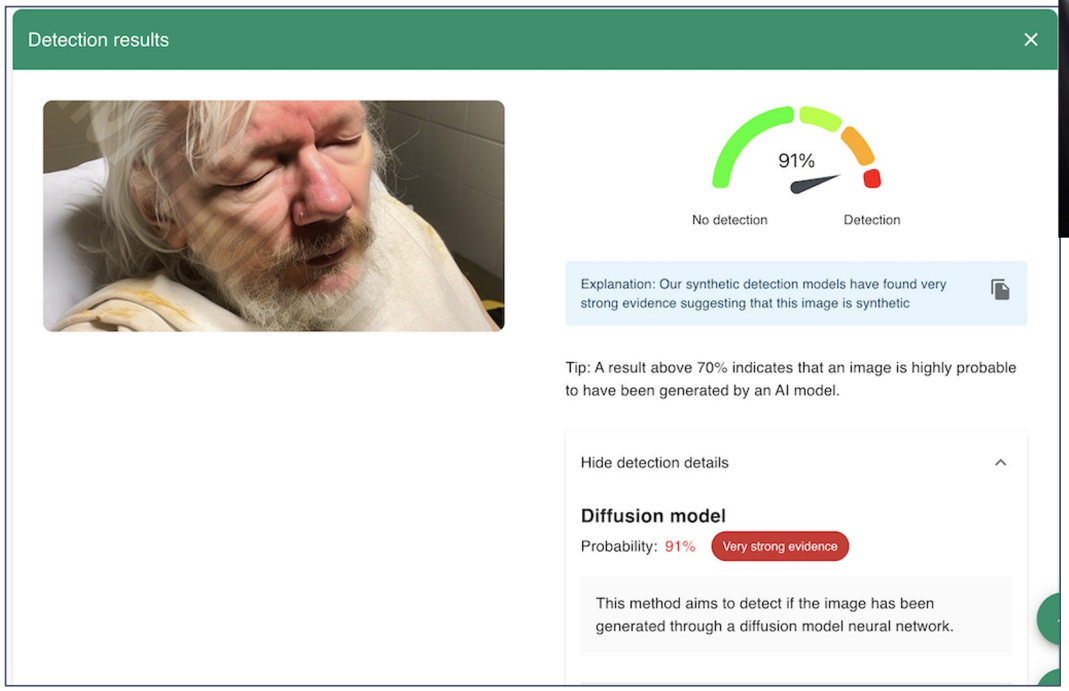
Overview

- Introduction to the Problem
- Challenges and Limitations
- vera.ai Research Contributions
- **Tools for Synthetic Media Detection in vera.ai**

vera.ai user facing tools

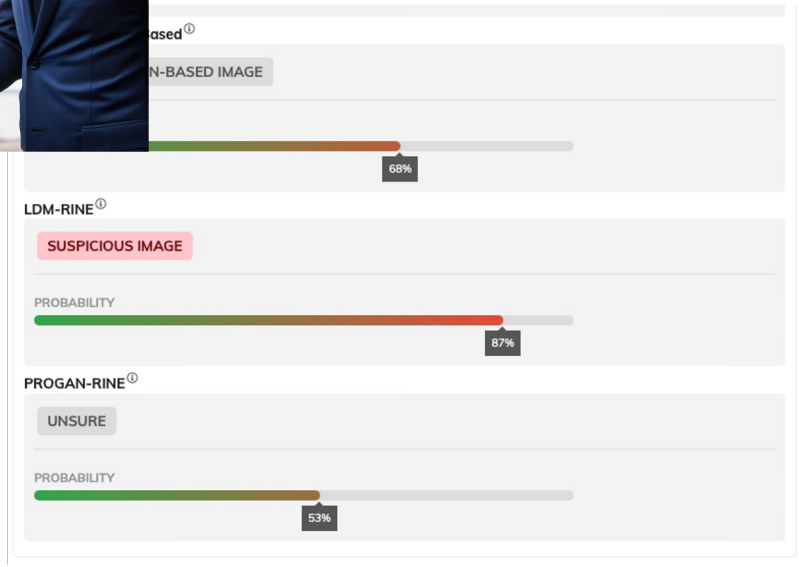
Verification Plugin

AFP



Truly Media

ATC
ATHENS TECHNOLOGY CENTER



Open Challenges

- **Integration Challenges**
  - Robustness and stability
  - Communicating AI results to journalists and fact-checkers
  - UI/UX trade-offs

- **Detection accuracy & reliability**
  - Constant adaptation to new generative models
  - Robustness against adversarial attacks and common post-processing operations

- **Fully synthetic video**
  - Ongoing work!

Luisa Verdoliva, Christos Koutlis /UNINA,CERTH-ITI

Contact: verdoliv@unina.it, ckoutlis.iti.gr

vera.ai

CERTH CENTRE FOR RESEARCH & TECHNOLOGY HELLAS

The University Of Sheffield.

ATC ATHENS TECHNOLOGY CENTER

KINIT

ontotext

Deutsche Welle

UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II

1506 UNIVERSITÀ DEGLI STUDI DI URBINO CARLO BO

AFP

UNIVERSITEIT VAN AMSTERDAM

école normale supérieure paris—saclay

université PARIS-SACLAY

EBU OPERATING EUROVISION AND EURORADIO

EU DISINFO LAB

Fraunhofer IDMT

Follow us on Twitter: @veraai_eu

Website: https://www.veraai.eu/

Co-financed by the European Union, Horizon Europe programme, Grant Agreement No 101070093.

Additional funding from Innovate UK grant No 10039055 and the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract No 22.00245