DETECTING SYNTHETIC IMAGES

KEY FINDINGS

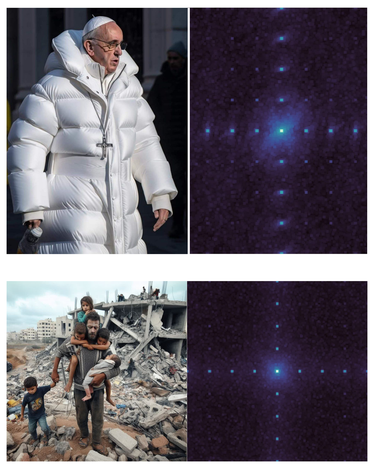
The exploration of five datasets containing real-world copies of synthetically generated images shared on social media platforms (one on the Pope in white coat and four Gaza-war related images) confirmed our initial hypothesis that **detection rate decreases over time**. It also showed the complementarity of our current methods [3, 4, 5]

More surprisingly, we found that image compression is not as much an issue as expected, both when extracted from the metadata and analysed on the image [2]. Remaining issues for research: exploring compression history and effects of re-sampling.

HOW DOES DETECTION WORK?

The GRIP-UNINA methods are detecting the traces through a ML classifier trained with fake images and their forensic traces. It returns a confidence score that the image was generated by AI-models.

Image generators are leaving invisible spectral artifacts as seen as bright spots in the visualisation.



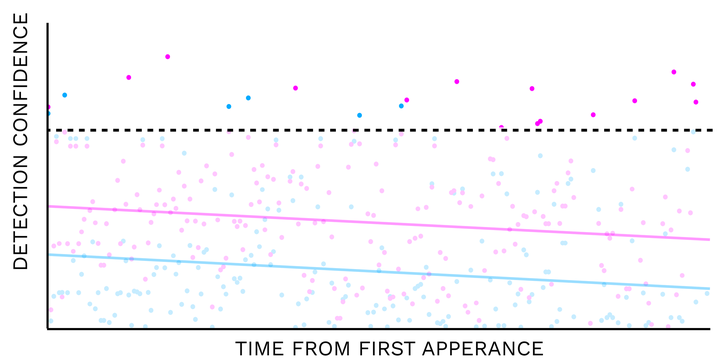
Both results were obtained using the "Synthbuster" method [1] developed by Quentin Bammey.

DATA COLLECTION

We collected fact-checked images generated by AI of the Pope and of the Gaza conflict using - as a first step - the Google Fact-Check Explorer. Then, through similarity search we collected 1000 different links per image which included images mostly from social media posts.

We automated the retrieval of the image urls and manually verified the image quality. screenshots of Twitter posts, cropped images, memes, image collages, and edited images were removed from the datasets .

POPE DATASET

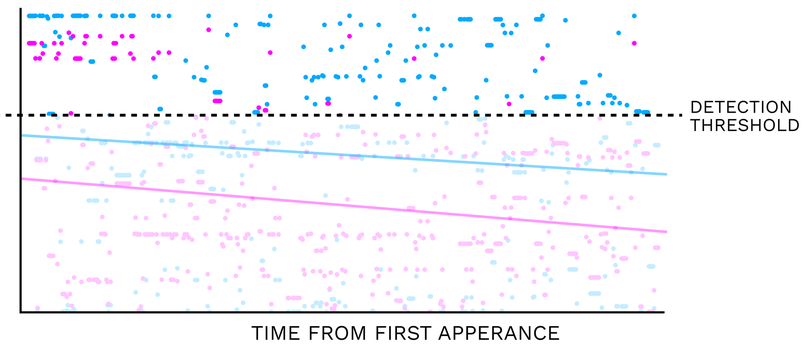


This scattered plot presents all the detection results performed by our tool methods on more 600 versions of the Pope generated hyperrealistic image.

The dotted black horizontal line represents the empirical threshold for detection at 70%.

A lot of the degraded copies are still not detected despite some promising results.

GAZA DATASET



We can observe on this chart that we detect more copies than on the chart of the left. However, the downward sloping trend lines on both charts are a sign that detecting the copies of the images is increasingly harder when the copies are farther apart in time.

WORKFLOW AND METHODOLOGY

Retrieval of AI-generated images on Google Fact Check Explorer

↓

Reverse Image Search

↓

1000 URLs for each of 4 images

↓

Crawl URLs for direct Image URLs

↓

Data Cleaning and Filtering

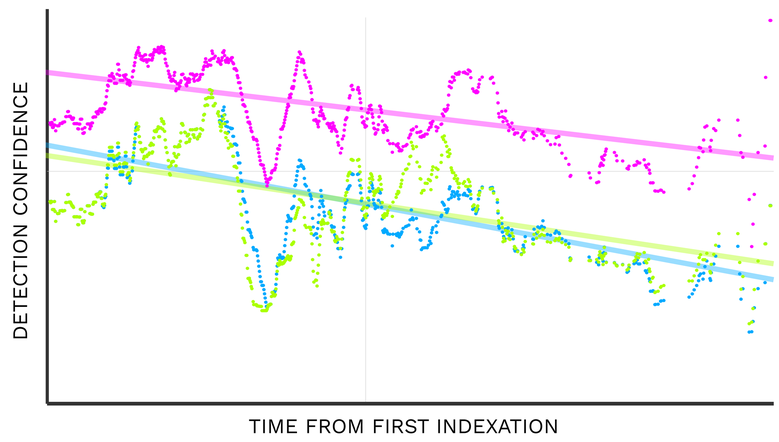
↓

Removed:
- Image Hashing and duplicates
- False Positives
- Mosaics, collages and watermarked images
- Video Thumbnails
- Paywalled websites

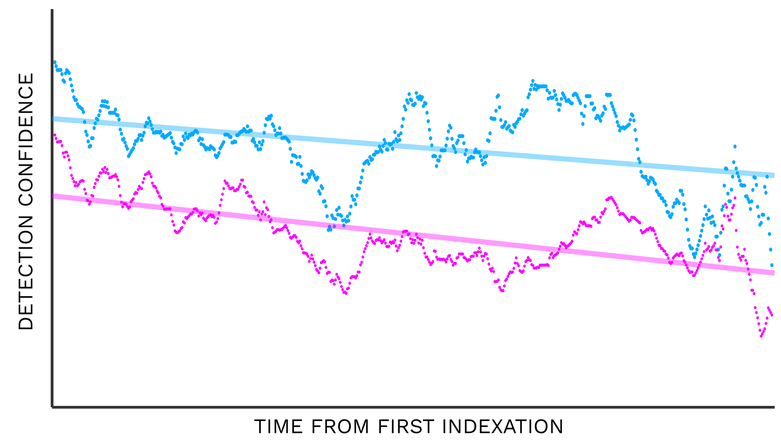
↓

Data Augmentation

- Google Sheet Formulas → Domain extraction
- Python → Resolution Compression rate
- Vera.ai API → Latent Diffusion Model Detection (LDM); Adaptive Diffusion Model Detection (ADM)

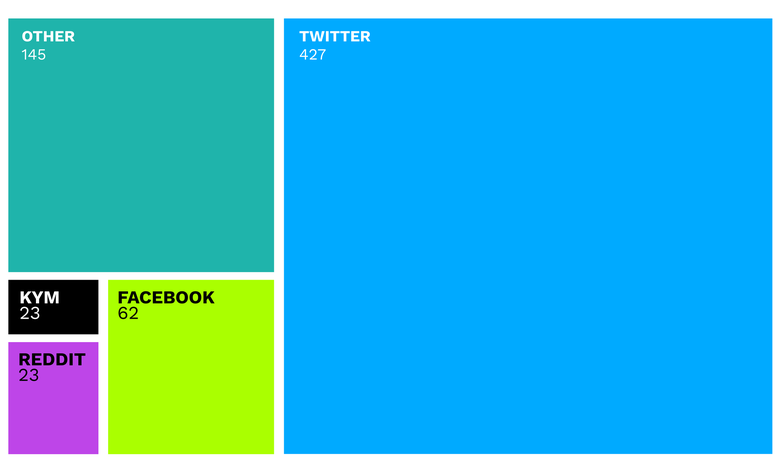


This chart shows the detection results of our two ML methods (new version of the LDM model in blue and paper version of LDM in Green, ADM model in pink) per indexation time. We observe a decrease of detection over time which confirms our preliminary hypothesis. Images posted on social networks are often degraded copies of the previous images.

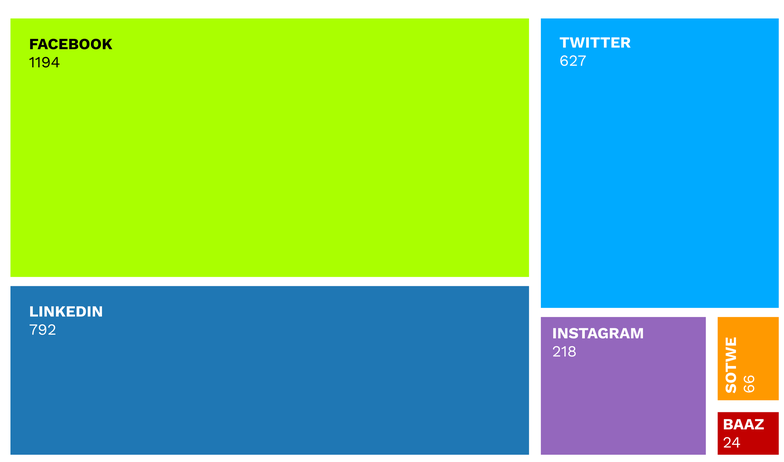


In this chart, the LDM model in blue performs better than the ADM model in pink. This result shows the complementarity of the two methods as sometimes ADM can perform better as displayed in the chart on the left.

We used a moving average to smooth the detection confidence curves in both charts to better identify the downward trend over time.



The pope images in our dataset mostly came from Twitter posts and mentions. This image from March 2023 has marked the rise of viral generative images on social networks. Our dataset focuses on the first images indexed by search engines to try to capture the first occurrences of the generated image.



This chart compiles the data from our Gaza dataset for 3 different pictures. Eight months after the first viral generated images, the spread is expanding to more social networks.

REFERENCES

1. Bammey, Q. (2024). Synthbuster: Towards Detection of Diffusion Model Generated Images. In IEEE OJSP.
2. Bammey, Q. (2024). JPEG Compression Quality Analysis for Image Forensics and AI-Generated Content Detection.
3. Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., & Verdoliva, L. (2023). On the detection of synthetic images generated by diffusion models. In IEEE ICASSP.
4. Cozzolino, D., Poggi, G., Corvi, R., Nießner, M., & Verdoliva, L. (2023). Raising the Bar of AI-generated Image Detection with CLIP.
5. Dogoulis, P., Kordopatis-Zilos, G., Kompatsiaris, I., & Papadopoulos, S. (2023). Improving Synthetically Generated Image Detection in Cross-Concept Settings. In Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation.

FACILITATORS

Denis Teyssou AFP
Kamila Koronska UvA
Richard Rogers UvA
Luca Draisci DensityDesign

PARTICIPANTS

Quentin Bammey
Bertrand Goupil
Dimitrios Karageorgiou
Matus Solcany
Sara M. Hammerschmidt
Kwan Suppaiboonsuk
Franziska Tietze
Frieder Uhlig
Youri van der Weide
Valentin Porcellini

AFP

digital methods initiative

University of Amsterdam

vera.ai
https://veraai.eu/

Vera.ai Website