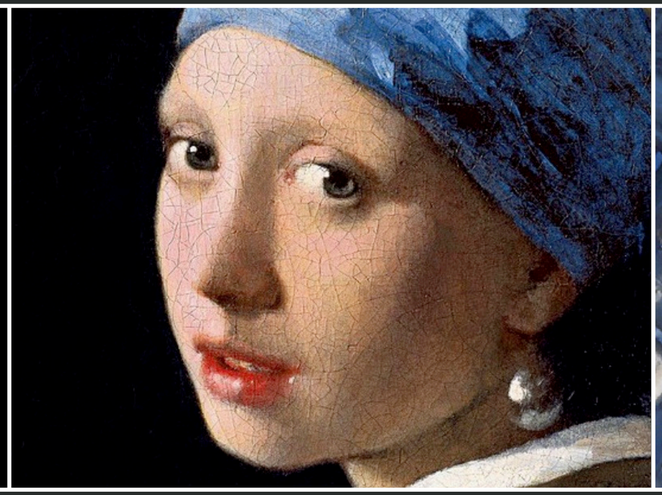

Source: Planet Labs, Inc.


Source: Zakharov et al.



DEEPFAKE GEOSPATIAL INFORMATION

CITIZEN MONITORING IN THE ERA OF SYNTHETIC MEDIA

Alex Glaser

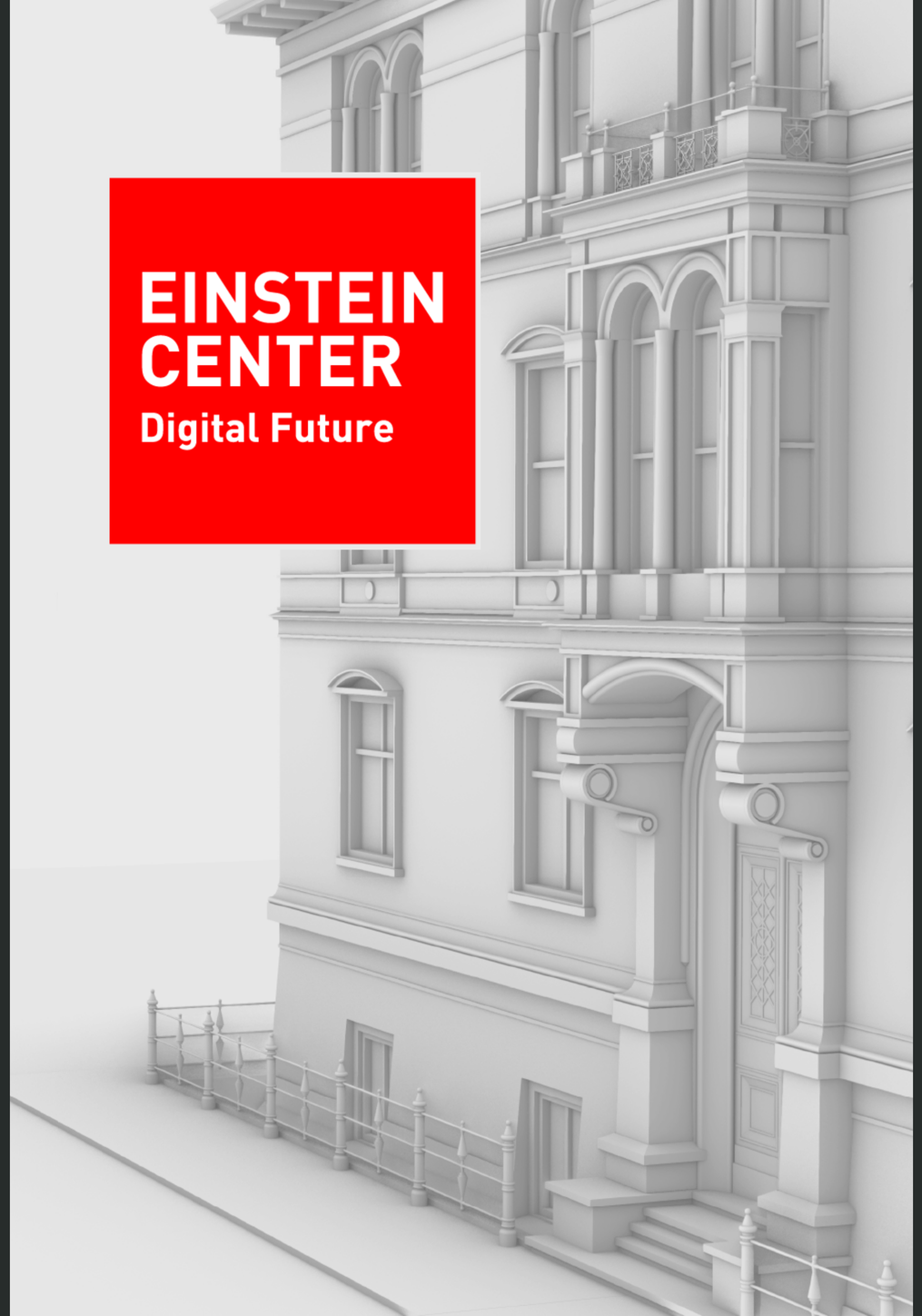
Program on Science and Global Security

Emerging Technologies Race, Nuclear Weapons, and Global Security
Princeton University, June 14–16, 2023

Revision 3b

BHT

Berliner Hochschule
für Technik

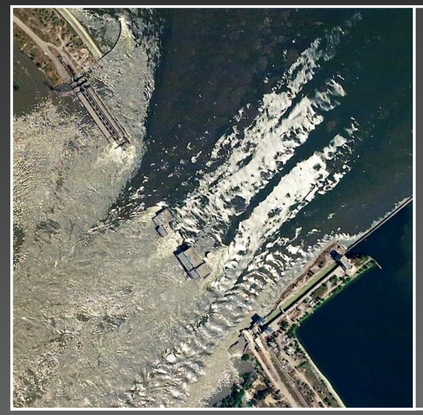


EINSTEIN
CENTER
Digital Future

SCIENCE &
GLOBAL SECURITY
PRINCETON UNIVERSITY

independent

TWO MAJOR DEVELOPMENTS



ABILITY TO MONITOR THE PLANET IN NEAR REAL-TIME

Evolving “megaconstellations” of optical imaging (and other) satellites with revisit times as short as 20 minutes; even high-resolution imagery becoming commercially available at scale

Relevant for many communities, including for “open-source intelligence” (OSINT) analysts

ABILITY TO GENERATE SYNTHETIC MEDIA THAT ARE INDISTINGUISHABLE FROM REAL MEDIA

With the advent of Generative AI (such as Stable Diffusion or DALL·E 2), it is becoming easier to generate realistic synthetic media and deepfakes — posing a range of challenges for society and policy

Dilemma to avoid: “When everything is possible, nothing really matters”

Source: Planet Labs (top) and Pablo Xavier, www.reddit.com/r/midjourney (bottom)

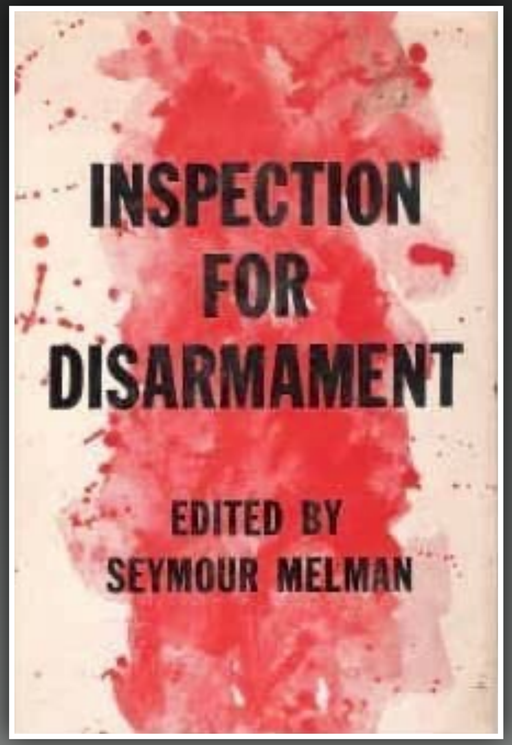
“Historically, it will turn out that there was this weird time when people just assumed that photography and videography were true. And now that very short little period is fading.”

Alexei A. Efros

November 2018

BACKGROUND

“INSPECTION BY THE PEOPLE”



From this viewpoint the problem may be posed: How can the manpower requirements for a major clandestine production effort be used to strengthen the possibilities of inspection for disarmament?

Inspection by the people is a method that would serve this purpose. In addition to the specific monitoring activities of the inspectorate, it would be invaluable to have a randomly distributed network of inspection that is based upon public support for inspection for disarmament. Such public support could reinforce the work of the inspectorate and could help to undercut evasion efforts that require substantial organizations and widespread production systems. The operation of effective world-wide inspection by the people would be facilitated if the disarmament agreements included provisions which made it a duty, an explicit obligation, of the citizens of participating countries to report violations to the international inspectorate.

Seymour Melman (ed.), *Inspection for Disarmament,* Columbia University Press, New York, 1958
see in particular: “Inspection by the People: Mobilization of Public Support” (pp. 38–44)

For a similar discussion, see Jerome B. Wiesner, ”Inspection for Disarmament,” Chapter 4 in *Arms Control: Issues for the Public,* Prentice-Hall, 1961

The Economist

What if bitcoin fell to zero?
Inside Xinjiang's economy
How to solve the chip shortage
Predicting pathogens

AUGUST 7TH–13TH 2021

The people's panopticon

Open-source intelligence comes of age

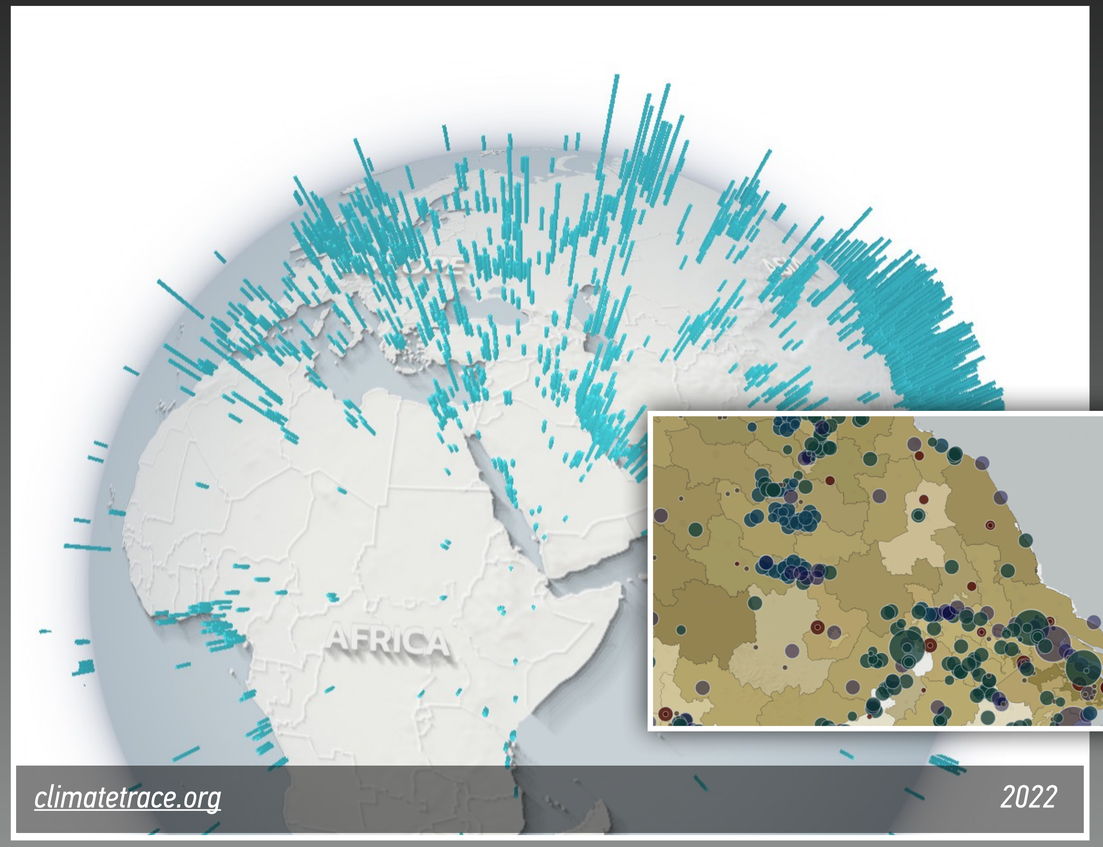
Trainspotting, with nukes

Geo4Nonpro, a crowdsourced project which let budding hobbyists and seasoned experts collaborate to annotate satellite pictures of everything from uranium mines in India to chemical-weapon facilities in Syria. "It's fun," says Mr Eveleth.

www.economist.com/weeklyedition/2021-08-07

ENVIRONMENTAL MONITORING

Skybox Imaging, www.youtube.com/watch?v=fCrB1t8MncY *2013*



climatetrace.org *2022*

spectrum.ieee.org/how-to-track-the-emissions-of-every-power-plant-on-the-planet-from-space

ARCHAEOLOGICAL SITE MONITORING

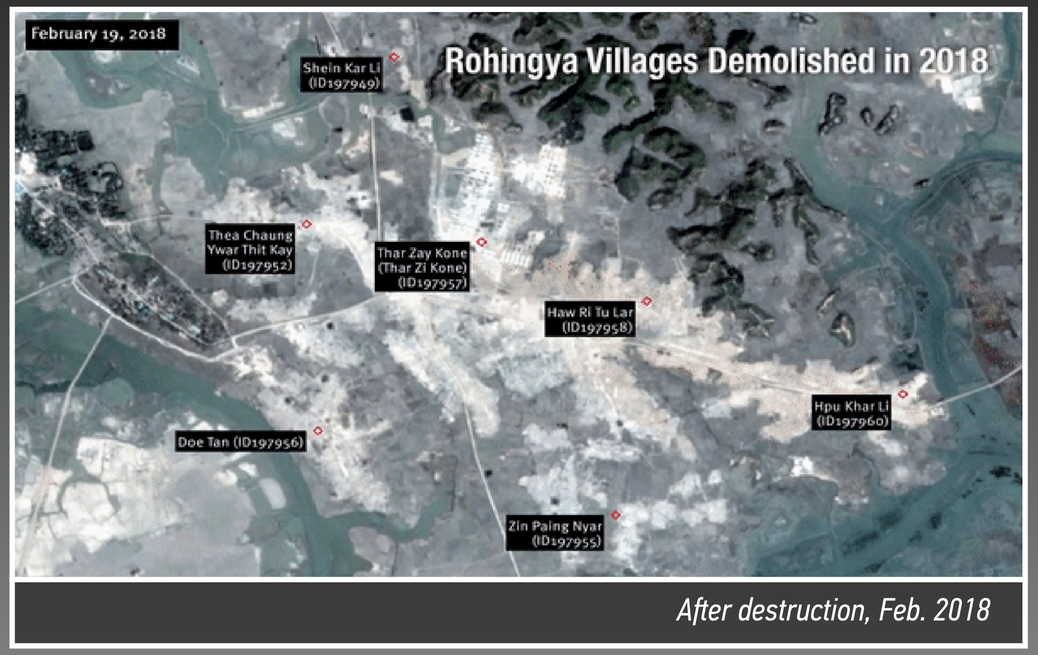
Ruins of Mari, Eastern Syria (34.551, 40.889) *September 2012*

Evidence of damage and looting *November 2014*

Jesse Casana and Elise Jakoby Laugier, “Satellite Imagery-based Monitoring of Archaeological Site Damage in the Syrian Civil War” *PLOS One,* 12 (11), November 30, 2017, doi.org/10.1371/journal.pone.0188589

HUMAN RIGHTS MONITORING

Myanmar (20.900, 92.366) *Prior to destruction, Jan. 2018*



After destruction, Feb. 2018

Burma: Scores of Rohingya Villages Bulldozed, New Satellite Images Show Destruction Indicating Obstruction of Justice, February 2018
www.hrw.org/news/2018/02/23/burma-scores-rohingya-villages-bulldozed and www.hrw.org/tag/rohingya

ICBM silo field, under construction; Copernicus Sentinel Data, January 2, 2023 (42.273 N, 92.682 E)
fas.org/blogs/security/2021/07/china-is-building-a-second-nuclear-missile-silo-field/

20 km (~ 12 miles)

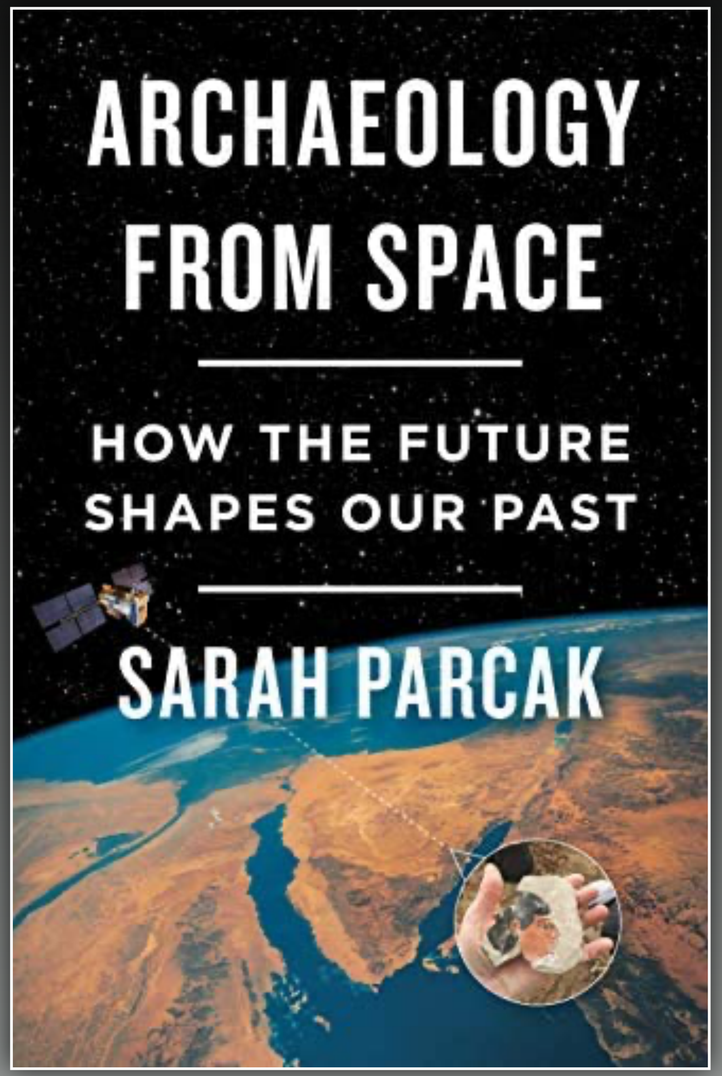
ISSUES & CHALLENGES

LACK OF ACCESS TO IMAGERY

"Analyzing the planet at scale with satellite imagery and machine learning is a dream that has been constantly hindered by the cost of difficult-to-access highly-representative high-resolution imagery."

Julien Cornebise, Ivan Oršolić, and Freddie Kalaitzis, Open High-Resolution Satellite Imagery: The WorldStrat Dataset — With Application to Super-Resolution, July 2022, arxiv.org/abs/2207.06418

citizenevidence.org/2020/07/06/using-artificial-intelligence-to-scale-up-human-rights-research-a-case-study-on-darfur/



The New York Times

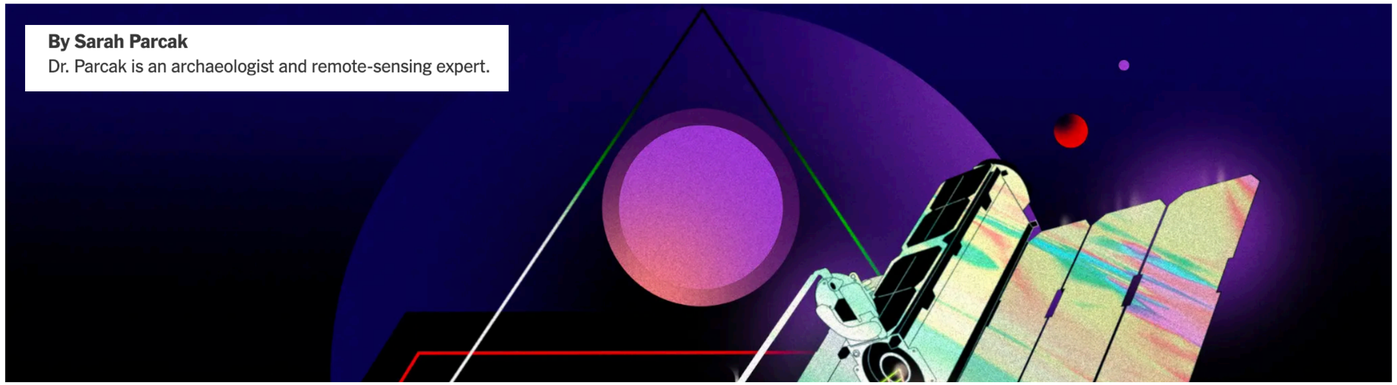
Opinion | THE PRIVACY PROJECT

Are We Ready for Satellites That See Our Every Move?

We should consider the ethical implications of satellites that can identify us, and our license plates, from space.

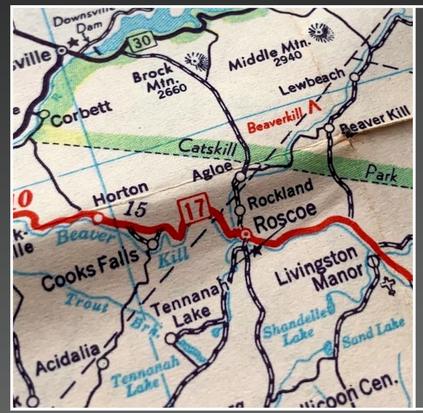
Oct. 15, 2019 4 MIN READ

By Sarah Parcak
Dr. Parcak is an archaeologist and remote-sensing expert.



www.nytimes.com/2019/10/15/opinion/satellite-image-surveillance-that-could-see-you-and-your-coffee-mug.html

mis GEOSPATIAL INFORMATION



GEOSPATIAL MISINFORMATION (THEN)

An old problem; fake locations and other inaccuracies have been part of mapmaking for centuries; including “copyright traps” and “paper towns” as a strategy to thwart plagiarism

Mark Monmonier, *How To Lie With Maps,* University of Chicago Press, 1996

GEOSPATIAL MISINFORMATION IN THE AGE OF AI

Few known examples, but circumstantial evidence suggests that AI has been used to manipulate scenes and pixels to create artifacts on satellite imagery for malicious purposes

Bo Zhao, Shaozeng Zhang, Chunxue Xu, Yifan Sun, and Chengbin Deng, “Deep Fake Geography? When Geospatial Data Encounter Artificial Intelligence,” *Cartography and Geographic Information Science,* 2021

Source: Esso Map, 1956 (top) and Planet Labs (bottom)

QUESTION 1

Can we generate & use synthetic satellite imagery to improve detection (or other) algorithms?

(when applied to real-world problems/imagery)

QUESTION 2

Can we use synthetic imagery
to assess the "true" potential of satellites
for monitoring & verification?

QUESTION 3

Can we help support efforts to confirm the authenticity of digital media?

(and, in particular, the provenance & authenticity of satellite imagery)

QUESTION 1

Can we generate & use synthetic satellite imagery to improve detection (or other) algorithms?

(when applied to real-world problems/imagery)

GENERATIVE ARTIFICIAL INTELLIGENCE

GENERATIVE ADVERSARIAL NETWORKS (~ 2015–2020)

Two neural networks compete with one another to make predictions
that are as accurate as possible (for example, distinguishing real from fake pictures)

Ian J. Goodfellow et al., Generative Adversarial Nets, arxiv.org/abs/1406.2661, June 2014
This Person Does Not Exist (StyleGAN, Nvidia, 2018)

FOUNDATION MODELS (2018–2023)

A large AI model is pre-trained on a vast quantity of unlabeled data resulting in a model
that can be adapted to a wide range of downstream tasks

Modalities include: text, code, imagery, music, video, … and many scientific applications

On Transformers, see: Ashish Vaswani et al., Attention Is All You Need, arxiv.org/abs/1706.03762, 2017

Source: this-person-does-not-exist.com (top) and Stable Diffusion (bottom)

SUBJECT-DRIVEN IMAGE GENERATION

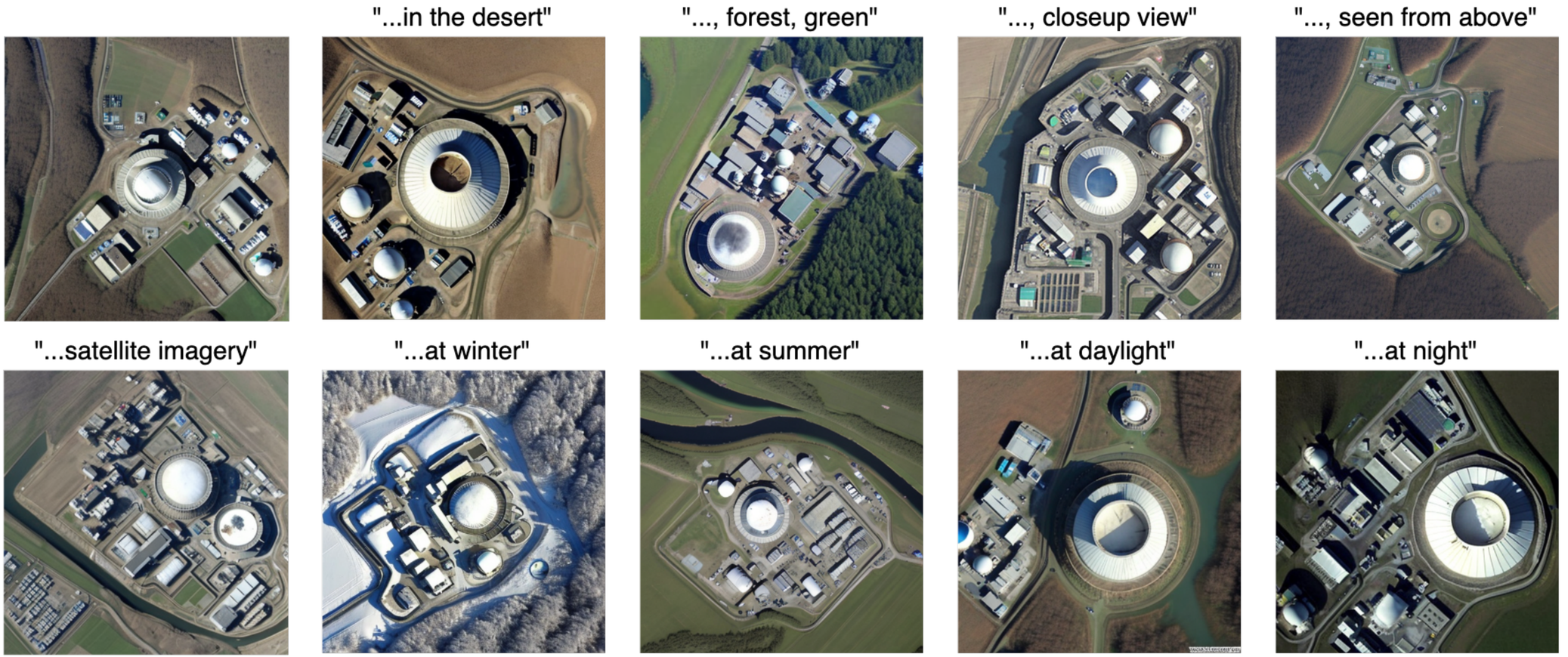
BY FINE-TUNING STATE-OF-THE-ART TEXT-TO-IMAGE MODELS (STABLE DIFFUSION, 2022)

Six variations of a single input image
of the Neckarwestheim nuclear power plant

Eight sample images of nuclear power plants from different regions of the world
(Overall, there are 202 input images of 185 unique plants in our dataset)

Vy Nguyen, *Machine Learning for Synthetic Satellite Images: Conditional Image Generation using a Vision-Language Model*
Master's Thesis, Berliner Hochschule für Technik, Berlin, May 15, 2023

“DREAMBOOTH NECKARWESTHEIM”

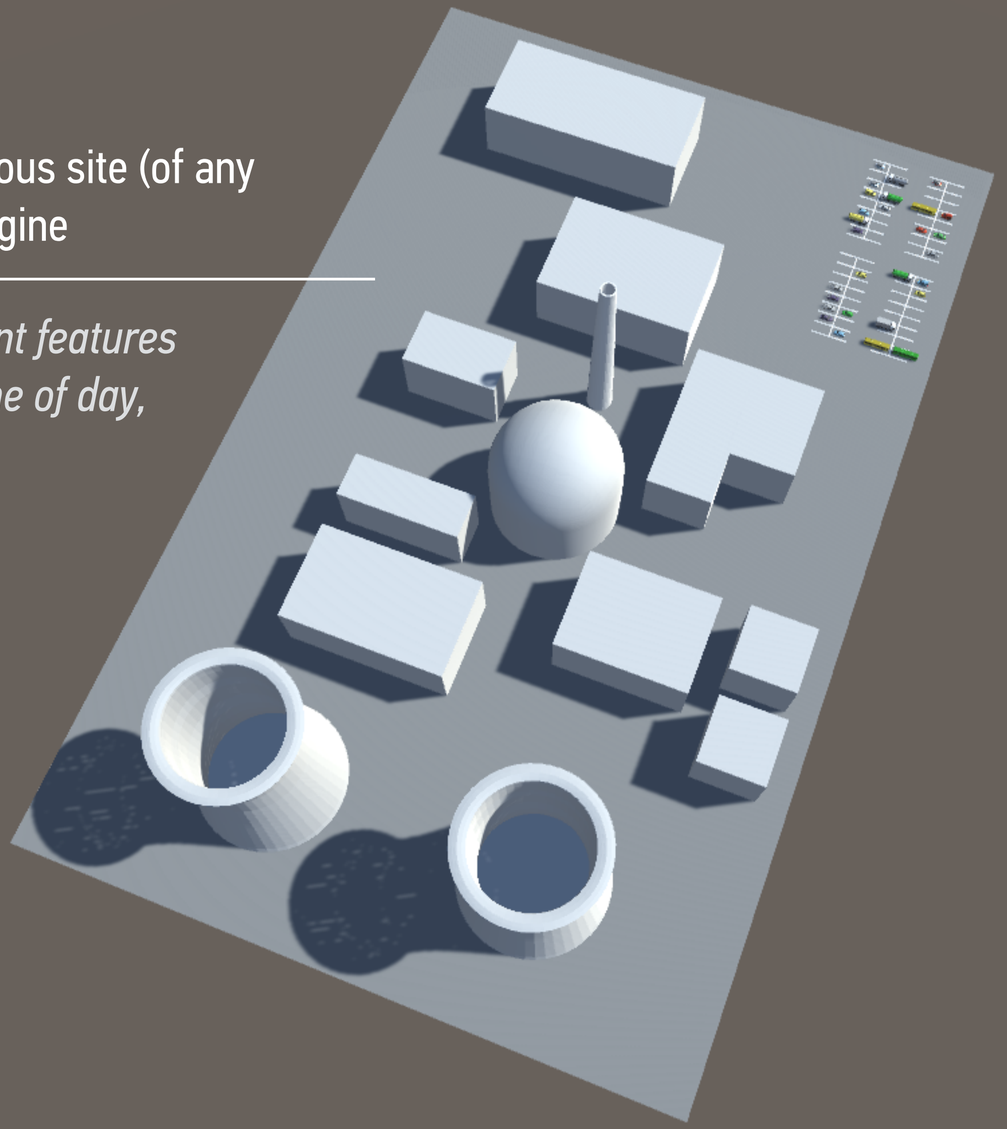


Vy Nguyen, *Machine Learning for Synthetic Satellite Images: Conditional Image Generation using a Vision-Language Model*
Master’s Thesis, Berliner Hochschule für Technik, Berlin, May 15, 2023

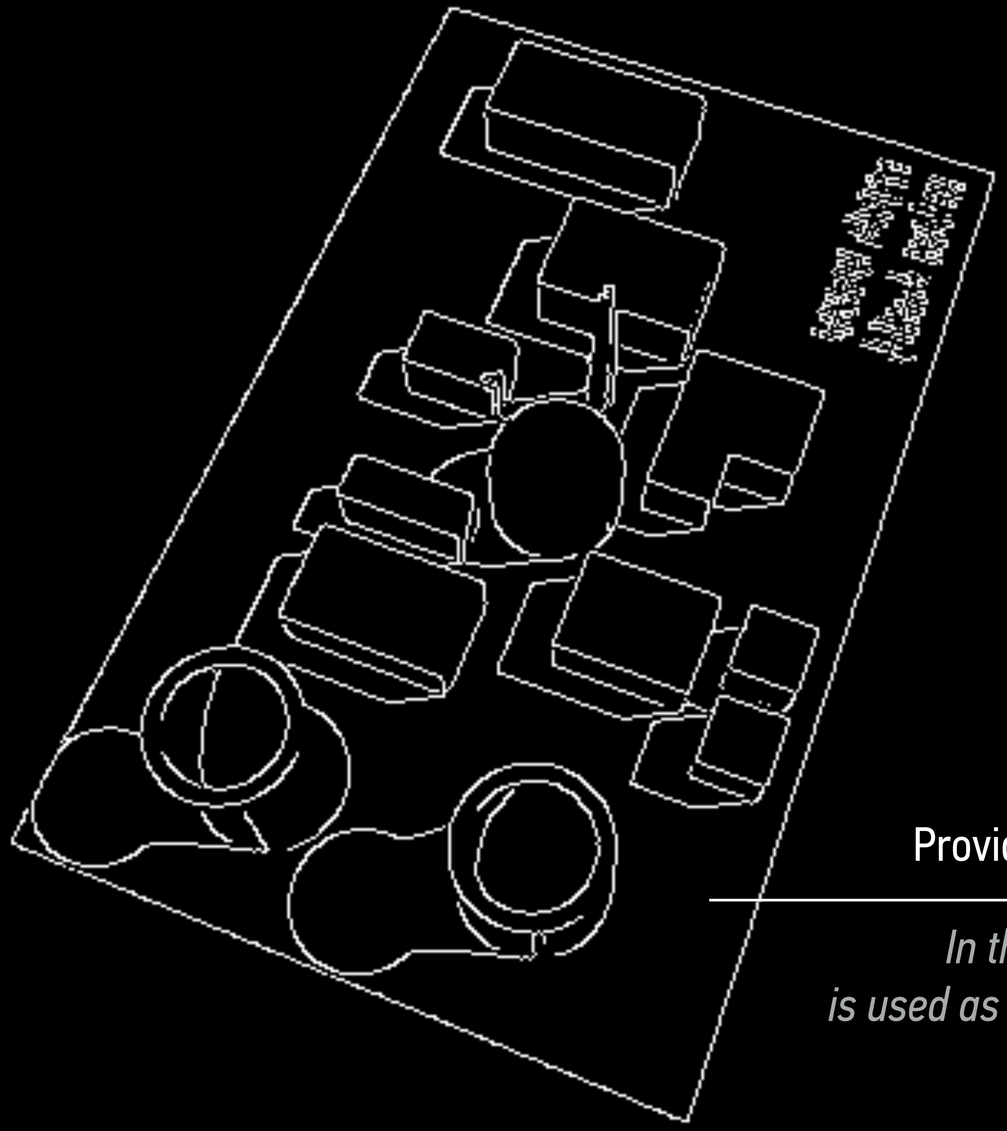
Procedurally generate layout of a fictitious site (of any desired type) using a modern Game Engine

---

Game Engine enables control of relevant features of scene, including: level of activity, time of day, cloud coverage, off-nadir angle, etc.



Courtesy: Johannes Hoster, BHT



Provide input modalities for structural guidance

In this example, the “canny edge” of the scene is used as an additional modality for a text-to-image composable adapter (“T2I CoAdapter”)

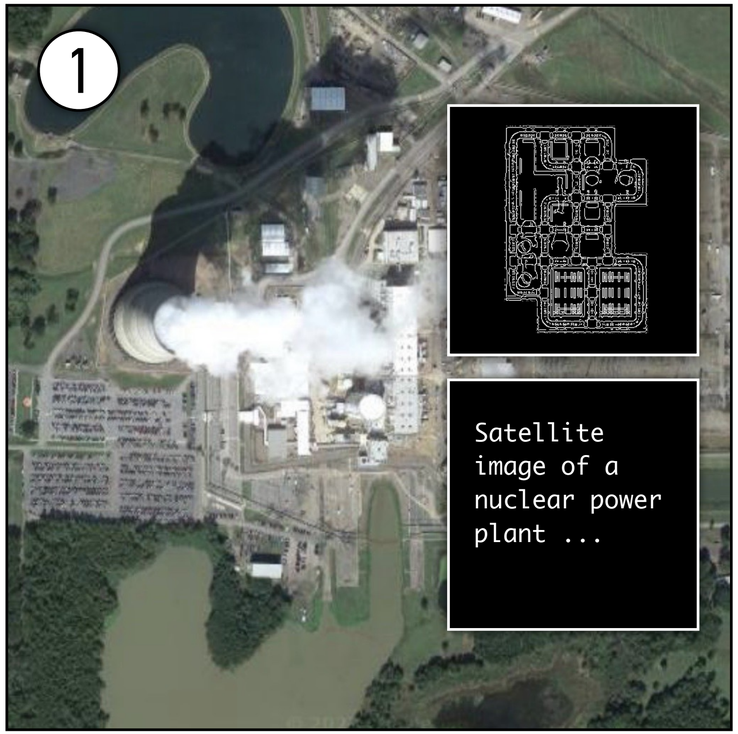
The canny edge complements the style image and the text prompt provided to the diffusion model

Courtesy: Johannes Hoster, BHT

Courtesy: Johannes Hoster, BHT

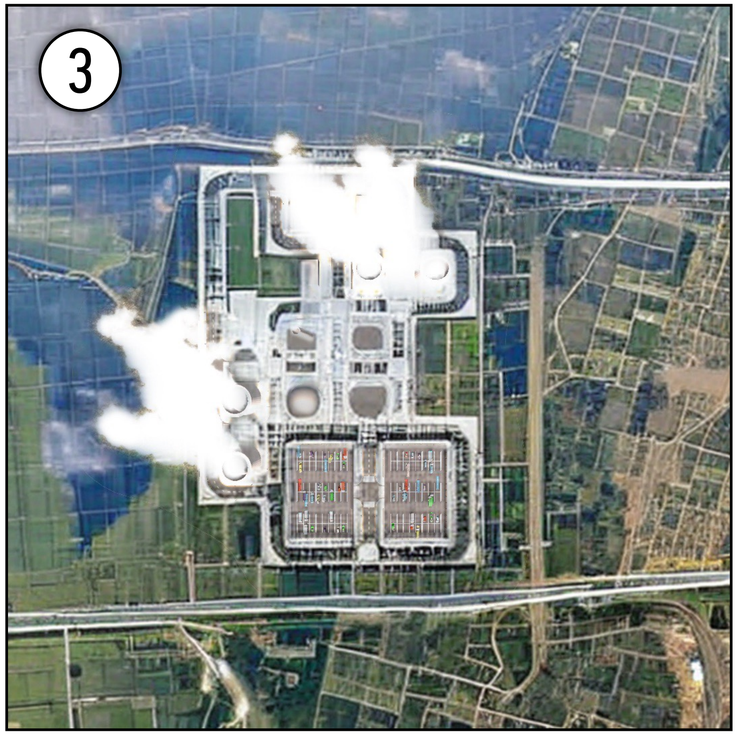
USING GAME ENGINES & MACHINE LEARNING

TO CREATE SYNTHETIC SATELLITE IMAGERY



Satellite imagery of real nuclear power plant

Synthesized image (with colormap of reference imagery)



Final image with details from game-engine render included

Johannes Hoster, Sara Al-Sayed, Felix Biessmann, Alexander Glaser, Kristian Hildebrand, and Vy Nguyen, *INMM & ESARDA Joint Annual Meeting,* Vienna, May 2023

QUESTION 2

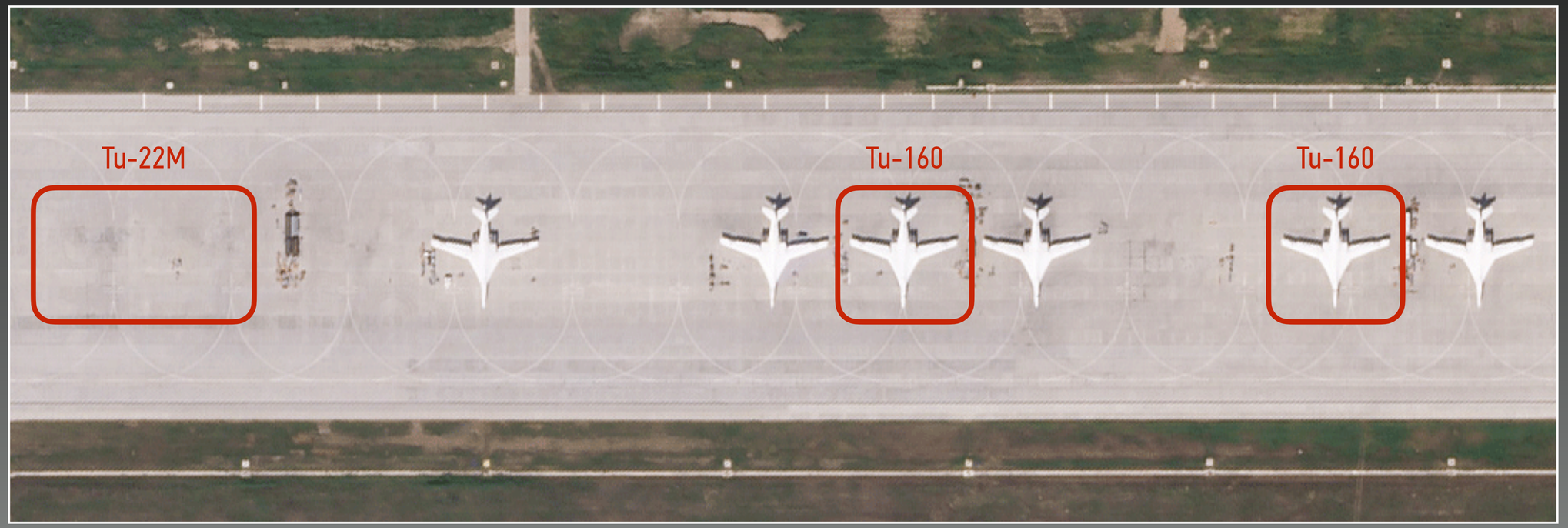
Can we use synthetic imagery
to assess the "true" potential of satellites
for monitoring & verification?

Fordow Enrichment Plant, Iran, in January 2016 (34.885 N, 50.996 E)
Iran's second enrichment plant was disclosed in September 2009; the plant itself is underground

Possible second uranium enrichment plant in North Korea near Pyongyang in January 2022 (38.957 N, 125.612 E)
Source: Google Earth

TOWARD "PUBLIC TECHNICAL MEANS"

NEAR REAL-TIME SATELLITE IMAGERY MAY ENABLE "PATTERN OF LIFE" ANALYSIS



Six images captured between August 18, 2018 and August 21, 2018 show the movement of the Tupolev Tu-22M (Backfire) and Tupolev Tu-160 (Blackjack) bombers on the flight line of Engels Air Base, Russia

Sub-daily rapid revisit capability (for SkySat, up to 12 times per day; global average of 7 times per day) may allow "pattern of life" analysis

www.planet.com/pulse/what-is-rapid-revisit-and-why-does-it-matter and www.planet.com/pulse/12x-rapid-revisit-announcement

QUESTION 3

Can we help support efforts to confirm the authenticity of digital media?

(and, in particular, the provenance & authenticity of satellite imagery)

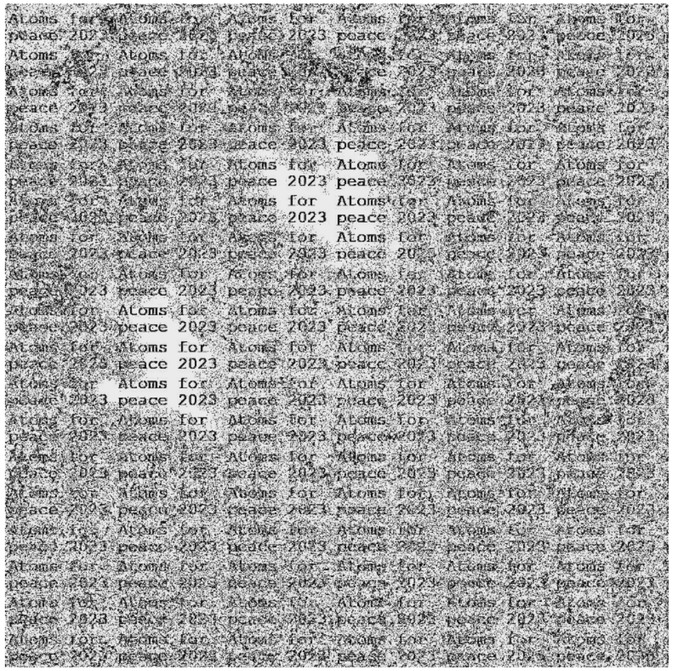
Corollary: Develop guidance and recommendations to ensure that the full potential of citizen-based monitoring can be realized

WATERMARKING SYNTHETIC MEDIA IS “EASY”

BUT IT DOES NOT REALLY ADDRESS (SOME) KEY CONCERNS ABOUT MISINFORMATION

Image with invisible watermark

Photograph Pixels	Red: 138 Even = Black	Red: 121 Odd = White	Red: 106 Even = Black	Red: 166 Even = Black	Red: 155 Odd = White
Invisible Watermark Pixels					



Retrieved watermark
“Atoms for peace 2023”

Source: invisiblewatermark.net (courtesy Johannes Hoster)

DIGITAL CONTENT PROVENANCE & AUTHENTICITY

WHAT TO WATERMARK: SYNTHETIC AND/OR AUTHENTIC MEDIA?

Ideally, watermark all authentic media; harder for some types of media than for others

Some industry efforts underway

- Coalition for Content Provenance and Authenticity (C2PA, c2pa.org)
  Led by Adobe; members include Microsoft, Intel, Arm, but also Canon, Nikon, and many others

SOME PRINCIPLES & CRITERIA FOR WATERMARKING OF DIGITAL MEDIA

- Security and robustness, i.e., watermarks that are resilient to manipulation
- Privacy, i.e., ability to control the privacy of information, including the identity of the source
- Scalability and flexibility, i.e., standards ought to be applicable to all common and future media types
- Universality and accessibility

See also: c2pa.org/principles

Source: www.natezeman.com (top) and Planet Labs (bottom)

CONCLUDING THOUGHTS

A NEW ERA OF GLOBAL TRANSPARENCY?

There is a widely shared expectation—or hope—that broad access to open-source information will enable the timely detection of non-compliance with relevant international agreements.

In reality, there are major obstacles to overcome to achieve this vision.

SYNTHETIC MEDIA ARE HERE TO STAY

Just like in the case of spam, malware, or phishing, "we should prepare ourselves for an equally protracted battle to defend against various forms of abuse perpetrated using generative AI." (Hany Farid, The Conversation, March 2023)

Source: Google Earth (top) and Chris Umé (bottom)

PROJECT TEAM & ACKNOWLEDGEMENTS

Vy Nguyen
Berliner Hochschule für Technik

Felix Biessmann
Berliner Hochschule für Technik

Rebecca D. Frank
University of Tennessee, Knoxville

Sara Al-Sayed
Princeton University

Igor Moric
Princeton University

Kristian Hildebrand
Berliner Hochschule für Technik

Alex Glaser
Princeton University

Johannes Hoster
Berliner Hochschule für Technik

Supported by the German Foundation for Peace Research (DSF)