Preventative studies should begin now for detecting AI-generated microscopy images

Jingshan S. Du* and Mingyi Zhang
Physical Sciences Division, Pacific Northwest National Laboratory, Richland, WA 99354, United States
*Email address: jingshan.du@pnnl.gov

ABSTRACT. With the rapid progress made in artificial intelligence (AI), text, images, and even videos that are often indistinguishable from human-created content can now be easily generated using new AI models. In this essay, we call attention to the potential misuse of advanced AI models to generate microscopy images, as it could challenge reviewers and fraud hunters in an unprecedented way. This concern is more than hypothetical because relevant open-sourced implementations and datasets are widely available. We urge that preventative studies on methods to detect microscopy image fabrication should begin now before potential misconduct materializes and becomes widespread.

Advanced artificial intelligence (AI) content creators such as large language model (LLM)-based chatbot ChatGPT and image generators DALL·E 2, Google Imagen, and Stable Diffusion have shocked the world with realistic outputs that are often indistinguishable from content generated by humans or actual photos. Worries recently emerged in academia about the potential of school assignments being completed using such tools; although some embraced it, several institutions chose to block access to ChatGPT on the school network[1]. In the research and scholarly publishing community, major publishers have issued statements to regulate and limit AI-generated text in manuscripts after ChatGPT was listed as a co-author on several journal articles and triggered extensive debates. New text detectors like GPTZero, DetectGPT, and an AI Text Classifier created by ChatGPT's own developers, OpenAI, have since been made available to flag possible paragraphs generated by LLM models.

In physical and life sciences research, various types of microscopy serve as compelling evidence of structures and processes of materials and biocomponents. For example, the successful synthesis of nanostructures is usually confirmed by their morphology and structural details observed through electron microscopy and scanning probe microscopy, and the effectiveness of new drugs delivered into cells can be evidenced by fluorophore-labeled imaging using confocal fluorescence microscopy. Until today, falsified micrographs are usually not too hard to identify under careful investigation because fabricating images entirely consistent with imaging theories and characteristics of real instruments is highly challenging and tedious. However, will AI-synthesized microscopy images soon be realistic enough to slip past the eyes of reviewers and fraud hunters?

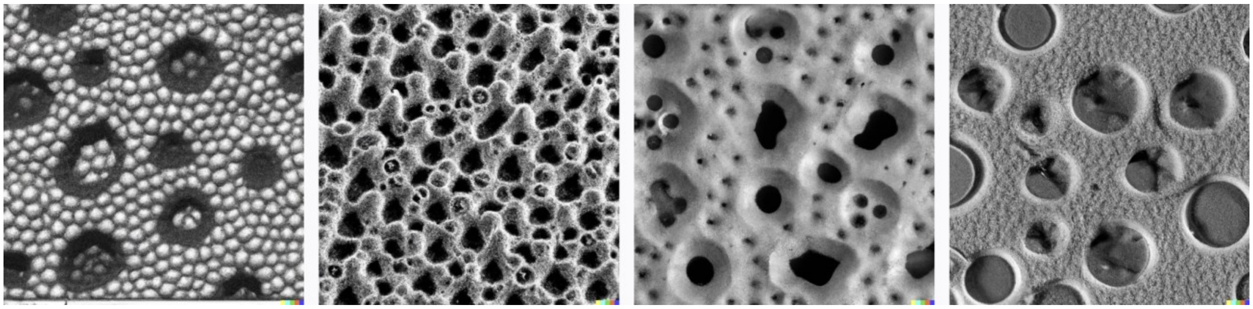

Figure 1 | DALL·E 2 can generate fake scanning electron microscopy images with complicated features from text prompts despite being trained on a general-purpose image dataset. Input prompt: "scanning electron microscopy image of an array of periodic holes on a flat surface." These images were generated on February 15th, 2023, at https://labs.openai.com/.

Today, general-purpose text-to-image converters such as the *diffusion*-based DALL·E 2[2] are already capable of synthesizing somewhat realistic micrographs with complicated features akin to nanotextured surfaces or biomineralization products (Figure 1). In this example, DALL·E 2 correctly captured two critical characteristics of scanning electron microscopy (SEM) images: (1) sharper, protruding features provide stronger signals, and (2) white noises distribute throughout the image. Suppose one properly trains a generative model using a dedicated experimental microscopy image dataset. In that case, paper mills or researchers with questionable intent could use such a program to rapidly produce scientifically realistic images of “samples” that do not exist. These images could be brand new, non-duplicated ones that are difficult to discover through manual inspection or existing image-checking software.

Given the widely available open-sourced implementations of such models and curated microscopy image datasets[3,4], it is perhaps only a matter of time before micrographic-realistic image generators come to life if they do not yet exist. Indeed, generative adversarial network (GAN)-based methods have already been reported for synthesizing certain types of confocal fluorescence micrographs[5] and SEM images[6,7] from specialized training datasets. With more advanced models like *transformer* and *diffusion* being employed, fake micrographs could soon begin to challenge the scientific community. In particular, some newer models leave very weak fingerprints in the spectral domain[8], and such features can also be mitigated through additional processing[9].

As AI technology rapidly advances and becomes easier to use, the emergence of new academic misconduct might be unavoidable. Preventative studies on methods to detect microscopy image fabrication should begin now before potential misconduct materializes and becomes widespread. Like how plagiarism has been combated through the integration of Crossref into existing workflows, it is never too late for funders, publishers, and institutions to invest in new integrity technologies and prepare for the future.

REFERENCES

1. The AI writing on the wall. *Nat. Mach. Intell.* **5**, 1–1 (2023).
2. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. Hierarchical Text-Conditional Image Generation with CLIP Latents. Preprint at https://doi.org/10.48550/arXiv.2204.06125 (2022).
3. Aversa, R., Modarres, M. H., Cozzini, S., Ciancio, R. & Chiusole, A. The first annotated set of scanning electron microscopy images for nanoscience. *Sci. Data* **5**, 180172 (2018).
4. Conrad, R. & Narayan, K. CEM500K, a large-scale heterogeneous unlabeled cellular electron microscopy image dataset for deep learning. *eLife* **10**, e65894 (2021).
5. Goldsborough, P., Pawlowski, N., Caicedo, J. C., Singh, S. & Carpenter, A. E. CytoGAN: Generative Modeling of Cell Images. Preprint at https://doi.org/10.1101/227645 (2017).
6. Du, H. & Shi, Z. Wafer SEM Image Generation with Conditional Generative Adversarial Network. *J. Phys. Conf. Ser.* **1486**, 022041 (2020).
7. Lambard, G., Yamazaki, K. & Demura, M. Generation of highly realistic microstructural images of alloys from limited data with a style-based generative adversarial network. *Sci. Rep.* **13**, 566 (2023).
8. Corvi, R. *et al.* On the detection of synthetic images generated by diffusion models. Preprint at https://doi.org/10.48550/arXiv.2211.00680 (2022).
9. Dong, C., Kumar, A. & Liu, E. Think Twice Before Detecting GAN-Generated Fake Images From Their Spectral Domain Imprints. in 7865–7874 (2022).

NOTES

The views expressed here do not necessarily reflect those of the authors’ employer or the US government.

ACKNOWLEDGEMENTS

J.S.D. acknowledges a Washington Research Foundation Postdoctoral Fellowship.

COMPETING INTERESTS

The authors declare no competing interests.