Patterns

CellPress
OPEN ACCESS

Opinion

Deepfakes: A new threat to image fabrication in scientific publications?

Liansheng Wang,[1] Lianyu Zhou,[1] Wenxian Yang,[2] and Rongshan Yu[1,2,3,*]
[1]Department of Computer Science, School of Informatics, Xiamen University, Xiamen, China
[2]Aginome Scientific, Xiamen, China
[3]National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China
*Correspondence: rsyu@xmu.edu.cn
https://doi.org/10.1016/j.patter.2022.100509

There is an increasing risk of people using advanced artificial intelligence, particularly the generative adversarial network (GAN), for scientific image manipulation for the purpose of publications. We demonstrated this possibility by using GAN to fabricate several different types of biomedical images and discuss possible ways for the detection and prevention of such scientific misconducts in research communities.

Data fabrication with the intention to manipulate research outputs is misconduct detrimental to high-quality scientific research. Among all forms of data fabrication, image fraud that manipulates images to distort their meanings is common as images play a central role in conveying research results in scientific publications. Inappropriate image manipulation thus poses a significant threat to trust within research communities as well as science's reputation with the general public. The scientific press continues to report a small but steady stream of cases of fraudulent image manipulation,[1] and yet more remain to be discovered with the development of large-scale automated image fraud detection technologies.[2]

In this opinion, we highlight the possibility of (mis)using the latest progress in artificial intelligence (AI) to perform image fabrication for the purpose of manipulating research outcomes. Such technologies are highly accessible nowadays due to the wide spread of their open-source implementations. We will also show that this kind of technology, if exploited inappropriately, can produce high-quality fraudulent images that could easily bypass existing image fraud detection technologies and fool even the most experienced researchers. It is thus timely to call our community's attention to this new threat to research integrity and reproducibility and the urgent need for developing effective and automated countermeasures to address this threat.

Scientific misconduct through image element reuse

Scientific misconduct can be damaging to trust in research communities as well as to the reputation of science in general public. A number of scientific misconduct cases have been found to make their fabricated images more difficult to be identified through inappropriate reuse of figure elements, typically with shifting, rotating, cropping, resizing, and/or making pixel-level modifications such as brightness, contrast, hue, and/or sharpness changes. Recently, Bik et al.[3] manually examined several thousand articles and found that 1.9% of them had some deliberate image manipulation. A larger scale of analysis on figure element reuse was performed by Acuna et al.[2] using an automated detection algorithm on a dataset comprising 760,000 open access articles and 2 million figures. After review by a three-person panel, around 0.6% of all articles were unanimously perceived as fraudulent, with inappropriate reuses occurring 43% across articles, 28% within an article, and 29% within a figure.

In response to this threat, many journals implemented image checking on their publications. Currently, most of them are done through random manual screening and relatively few have publisher-wide automated processes. However, with the development of automated image duplication detection methods as well as the increasing availability of shared databases of all published images across publishers, it can be envisioned that less room will be left for scientific misconduct through image element reuse in future once duplication detection becomes a routine practice for publishers to screen image elements in submitted manuscripts.

Deepfake as a new threat

Despite the endeavors in image manipulation identification on a large scale, emerging AI-based image synthesis technologies could render such efforts obsolete. One such technology is the generative adversarial network (GAN), which was introduced by Ian Goodfellow in 2014.[4] GANs are unsupervised generative models that learn an underlying distribution implicitly from a given set of training samples. Briefly, two competing neural networks are jointly trained in a GAN: a generative model *G* that captures the data distribution of the training samples and a discriminative model *D* that estimates the probability that a sample comes from the domain where the sample data were drawn rather than *G*. The training objective for *G* is to maximize the chance of *D* to make a mistake (Figure 1A).

GAN excels in generating highly realistic-looking synthetic contents and has gained significant success in many computer vision tasks. However, it can also lead to the advent of artificial misinformation if used inappropriately. In 2017, the DeepFake, a software based on GAN to create synthetic media in which a person

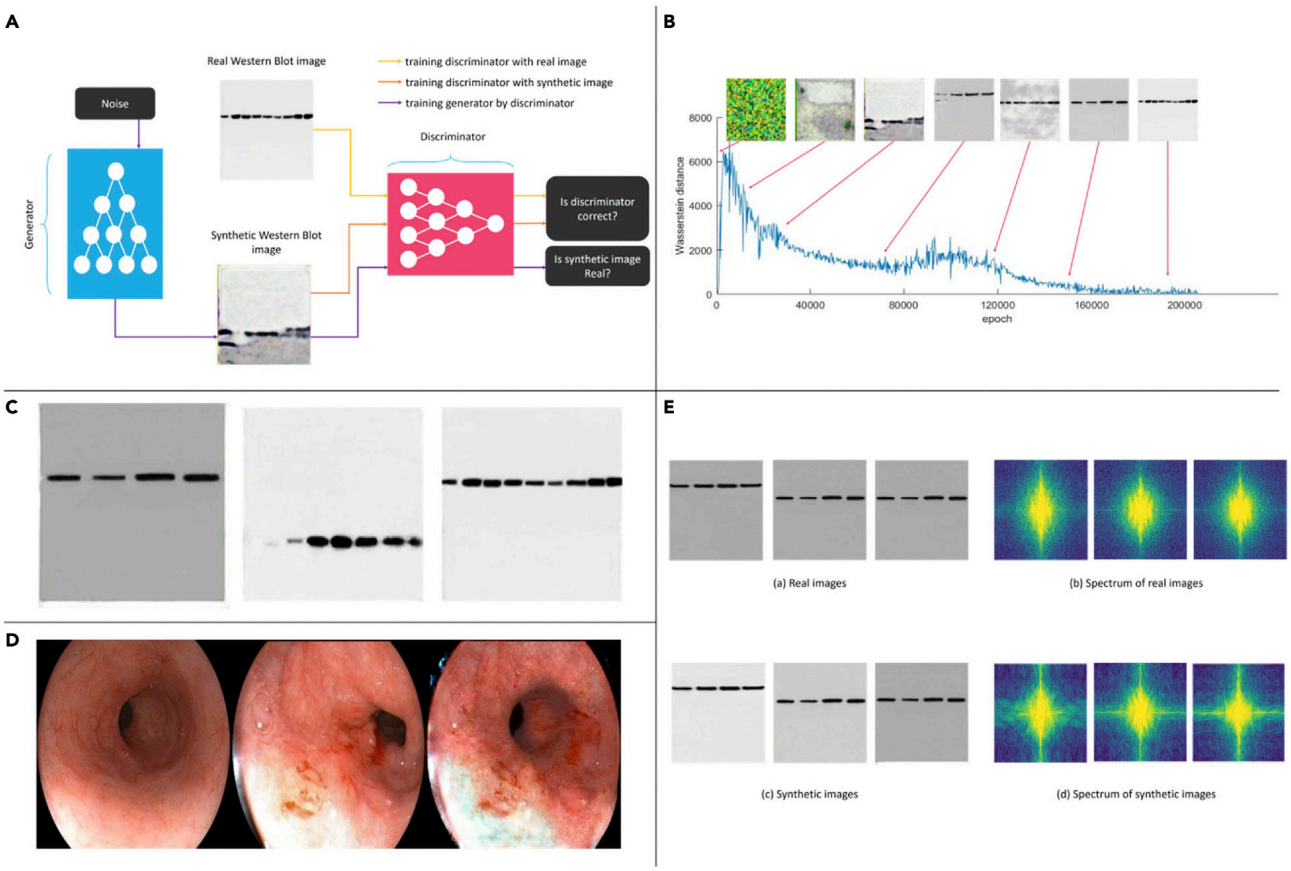

Figure 1. Workflow and example usage

(A) The GAN pipeline.

(B) The Wasserstein distance reduces when training epochs increase and the generated images at different training epochs.

(C) Examples of generated western blot images.

(D) Examples of generated esophageal cancer images.

(E) The synthetic images from GAN have more high-frequency parts than the real images.

in an existing image or video is replaced with someone else's likeness, appeared in Reddit. Despite most online platforms forbidding this application, it opened the Pandora's box that more ill-purposed software applications creating deceptive media appeared.

To demonstrate the potential impact that deepfake-like technologies can bring to our research community, we show two examples of using GAN to produce fake images that could potentially be used for scientific publication. Our demonstrations were constructed using well-known GAN network structures on a commodity personal computer with one Nvidia GTX 1080Ti GPU, which can be easily reproduced by people with a similar technical background.

Example 1. The western blot is widely used across a broad range of life science and clinical disciplines due to its simplicity and ability to clearly show the presence of specific proteins by size and the binding of an antibody. It also became a popular target for inappropriate image element duplication in scientific publications (https://scienceintegritydigest.com/2019/11/23/scanning-for-duplications/). To train a GAN that can synthesize western blot images, we retrieved 3,000 authentic western blot images using web crawlers as our training dataset. All images were resized to 256 × 256 pixels. We used a 15-layer network with two additional residual blocks as the discriminator and a 34-layer network as the generator. The training process was conducted until both networks converged. The intermediate results at different training epochs during the training process are shown in Figure 1B, and some examples of the synthetic western blot images after the networks are converged are shown in Figure 1C.

Example 2. In a more challenging example, we used CycleGAN[5] to artificially impose a specific condition, esophageal cancer, on gastroscope images from cancer-free locations of intestine. We collected 100 images from the MICCAI EndoVis Challenge (https://endovissub-barrett.grand-challenge.org/), including 50 positive images with esophageal cancer and 50 negative images from cancer-free locations. The target was to impose the desired feature naturally on the negative images to turn it into positive images. CycleGAN, which can transform images to a target data domain, provides a basic and efficient solution for such image-to-image translation tasks. Specifically, in each training iteration, the model takes a pair of images as input, including a negative

image A and a positive image B. The model then converts A into a positive one according to the features of image B, symmetrically turns image B into a negative one, and outputs this pair of images. An example of the synthetic images with the imposed condition generated with trained network is shown in Figure 1D.

How can we detect them?

It is very difficult, if not impossible, for the naked eye to distinguish between real and fake images generated by AI. In a subjective quality-evaluation test, we invited three board-certified biomedical specialists to evaluate the quality of the generated western blot images. A series of image groups were displayed to the subjects, where each group contained three real images and a fake one in random order unknown to the subjects. The subjects were then asked to identify the fake one. Among them, two experts achieved accuracy levels of 10% and 30% respectively, suggesting that they did not perform better than random guessing in identifying the fake western blot images generated from GAN. The third expert achieved an accuracy of 60% based on a subtle difference between real and fake western blot images; i.e., the boundary between synthetic blots and the background is not as smooth as that between real blots and the background. This is introduced by the intrinsic properties of GAN-based methods and can be dealt with by further optimizing the network model.[6]

Another possible solution is to examine the images in its frequency domain. It has been shown that synthetic images from GAN may have some common artifacts that are barely noticeable in the spatial domain but that become apparent once the images are converted to the frequency domain.[7] The most well-known artifact is the checkerboard effect due to the image upsampling operation in GAN.[8] A well-trained generator may gloss over this effect, making it hard to be discovered by the naked eye in the spatial domain. Yet the high-frequency artifacts in the spectrum cannot be completely removed, as exemplified in Figure 1E.

Since the fake images are not duplicated from existing images, traditional computational image duplication detection techniques will presumably fail on them. Fortunately, there exist other effective computational methods to discern the subtle differences between real and fake images. Due to limited data, network structure,[8] and optimization algorithms,[6] the GANs can hardly fit the distribution of real data. In particular, it has been shown that the current GAN-based image synthesis methods have common defects that can be leveraged to develop generic classifiers to identify synthetic images from different GAN generators.[7] To validate this concept, we built a classifier using the convolutional neural network (CNN) to identify synthetic western blot images generated from our previous tests and achieved an area under the operational curve (AUC) of 0.8542 (95% confidence interval [CI]: 0.8379–0.8705) on a cohort of 1,000 real images and 1,000 synthetic images.

Although our trials suggest that there exist effective methods to discern GAN generated images, it does not guarantee that people can be relieved. In the 2020 Deepfake Detection Challenge (https://www.kaggle.com/c/deepfake-detection-challenge), the champion algorithm only achieved a score of 0.42798, evaluated by log loss, indicating that a significant portion of synthetic images could not be detected. In addition, since the effectiveness of most of the current fraudulent detection methods relies on the flaws of GAN, these methods may quickly become obsolete with the fast evolution of deep learning technologies.

Machine intelligence algorithms have provided us incredible capabilities in making data-driven decisions. With the maturation of the technology and the surge of large amounts of data, we have seen AI dipping into every industry from the demonstrated surgery performed by a robot to the rise in autonomous vehicles. The development of AI has also granted us the power to resolve complex scientific problems using data-driven approaches, with applications ranging from the discovery of new genetic targets of diseases to the fabrication of new chemical materials. Unfortunately, without proper guidelines in place, AI could also create new threats to many aspects of our society due to its tremendous capacity for evil when being misused. Deepfake is a good example of the misuse of AI technologies. In this opinion, we have shown that deepfake can be leveraged as a new tool for image fabrication and that the resulting images could easily circumvent existing deterring measurements when they are used for publications. Therefore, it is necessary to develop and implement preemptive measures from both policy and technology perspectives before it starts to hurt science's integrity and reputation.

DECLARATION OF INTERESTS

R.Y. and W.Y. are shareholders of Aginome Scientific.

WEB RESOURCES

2020 Deepfake Detection Challenge, https://www.kaggle.com/c/deepfake-detection-challenge

MICCAI EndoVis Challenge, https://endovissub-barrett.grand-challenge.org/

Science Integrity Digest blog, https://scienceintegritydigest.com/2019/11/23/scanning-for-duplications/

REFERENCES

1. Cromey, D.W. (2012). Digital images are data: and should be treated as such. Methods Mol. Biol. *931*, 1–27. https://doi.org/10.1007/978-1-62703-056-4_1.
2. Acuna, D.E., Brookes, P.S., and Kording, K.P. (2018). Bioscience-scale automated detection of figure element reuse. bioRxiv. https://doi.org/10.1101/269415.
3. Bik, E.M., Casadevall, A., and Fang, F.C. (2016). The prevalence of inappropriate image duplication in biomedical research publications. MBio *7*. https://doi.org/10.1128/mbio.00809-16.
4. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, *27*, pp. 2672–2680.
5. Zhu, J.Y., Park, T., Isola, P., and Efros, A.A. (2017). Unpaired image-to-image translation using cycle- consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (IEEE), pp. 2242–2251. https://doi.org/10.1109/iccv.2017.244.
6. Mertikopoulos, P., Papadimitriou, C., and Piliouras, G. (2018). Cycles in adversarial regularized learning. In Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (SIAM), pp. 2703–2717.
7. Marra, F., Gragnaniello, D., Verdoliva, L., and Poggi, G. (2019). Do GANs leave artificial fingerprints? In Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) (IEEE), pp. 506–511. https://doi.org/10.1109/mipr.2019.00103.
8. Zhang, X., Karaman, S., and Chang, S.-F. (2019). Detecting and simulating artifacts in GAN fake images. In Proceedings of the 2019 IEEE International Workshop on Information Forensics and Security (WIFS) (IEEE), pp. 1–6.

https://doi.org/10.1109/WIFS47025.2019.9035107.

About the authors

Liansheng Wang received a PhD in computer science from the Chinese University of Hong Kong. He is currently an associate professor in the Department of Computer Science, Xiamen University, Xiamen, China. His research interests include medical image processing and analysis, machine learning, big medical data.

Lianyu zhou received a BS from Xiamen University in 2020 and now is a master's student in the Department of Computer Science, Xiamen University, Xiamen, China. His main research interests include medical image analysis and machine learning.

Wenxian Yang is the CTO of Aginome Scientific. Her research interests include signal processing, data analytics, and bioinformatics.

Rongshan Yu is currently with National Institute for Data Science in Health and Medicine, School of Informatics, Xiamen University, as a professor. His research interests include statistical signal processing and its applications in bioinformatics.