

LETTER TO THE EDITOR

Open Access



Visual Turing test is not sufficient to evaluate the performance of medical generative models

Shoichiro Yamamoto¹ and Akinori Higaki^{1*}

Keywords Artificial intelligence, Generative adversarial networks, Image processing (computer-assisted), Medical image synthesis, Visual Turing test

To the Editor,

We read with great interest the article by Wang et al. [1], reporting that generative adversarial networks (GANs) could generate synthetic ground glass opacities (GGOs) in computed tomography. While we appreciate their ambitious research to advance clinical radiology, we feel that the performance evaluation of the GANs is insufficient for their aim.

In their study, the authors stated that the model performance was evaluated by both subjective and objective approaches, namely the visual Turing test (VTT) and the distribution of radiomic features. We agree that VTT is a suitable approach to assess the realism of synthesized medical images [2], but a low VTT score does not guarantee the diversity of the generated data; it tells us they just look real. As the authors admitted as a limitation in the “Discussion” section, about 40% of the distributions of the radiomic features (e.g., NGTDM coarseness) were significantly different between generated and original images. Therefore, we suspect that their generative model may only be able to produce biased images due to the so-called mode collapse phenomenon [3]. If this were the

case, it would diminish the usefulness of the data augmentation for classification tasks.

It is true that there is no single universal metric to assess the model performance and the quality of generated data; therefore, we need to combine several indicators, such as inception score, Fréchet inception distance, and geometry score [4, 5]. In addition to these, the image quality can be also evaluated quantitatively by NIQE, PIQE, and BRISQUE scores, as Oyelade and colleagues have demonstrated for mammography images [6]. As a practical matter, the images presented in the article are so small in size and resolution that the readers cannot fully appreciate what kind of images the GAN model has produced.

In summary, we believe that the authors need to provide more example images of the generated GGO and evaluate their GAN in several other ways to ensure the quality of data synthesis.

Authors' contributions

AH conceptualized and drafted the manuscript. SY reviewed and revised the manuscript. The authors read and approved the final manuscript.

Funding

The authors declare that they received no external funding concerning this article.

Availability of data and materials

Not applicable.

Declarations

Ethics approval and consent to participate

This article is based on previously conducted studies and does not contain any studies with human participants or animals performed by the authors.

This comment refers to the article available online at <https://doi.org/10.1186/s41747-022-00311-y>.

*Correspondence:

Akinori Higaki
keroplant83@gmail.com

¹ Department of Cardiology, Pulmonology, Hypertension and Nephrology, Ehime University Graduate School of Medicine, 454 Shitsukawa, Toon, Ehime 791-0295, Japan

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 21 March 2023 Accepted: 21 April 2023

Published online: 10 July 2023

References

1. Wang Z, Zhang Z, Feng Y et al (2022) Generation of synthetic ground glass nodules using generative adversarial networks (GANs). *Eur Radiol Exp* 6:59. <https://doi.org/10.1186/s41747-022-00311-y>
2. Higaki A, Kawada Y, Hiasa G, Yamada T, Okayama H (2022) Using a visual Turing test to evaluate the realism of generative adversarial network (GAN)-based synthesized myocardial perfusion images. *Cureus*. 14:e30646. <https://doi.org/10.7759/cureus.30646>
3. Bau D, Zhu J-Y, Wulff J et al (2019) Seeing what a GAN cannot generate. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE; 2019. p4501–4510. <https://doi.org/10.1109/ICCV.2019.00460>
4. Shmelkov K, Schmid C, Alahari K (2018) How good is my GAN? Improving and optimizing operations: things that actually work - Plant Operators' Forum 2004:218–234. <https://doi.org/10.1007/978-3-030-01216-8>
5. Borji A (2019) Pros and cons of GAN evaluation measures. *Comput Vis Image Underst* 179:41–65. <https://doi.org/10.1016/j.cviu.2018.10.009>
6. Oyelade ON, Ezugwu AE, Almutairi MS, Saha AK, Abualigah L, Chiroma H (2022) A generative adversarial network for synthetization of regions of interest based on digital mammograms. *Sci Rep* 12:1–30. <https://doi.org/10.1038/s41598-022-09929-9>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

Submit your next manuscript at ▶ [springeropen.com](https://www.springeropen.com)
