Datasheets for GenImage Dataset

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang,
Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, Yunhe Wang
Huawei Noah's Ark Lab
{zhumingjian, yunhe.wang}@huawei.com

A Motivation

1. For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
GenImage is proposed for facilitaing the development of detectors for identifying the AI-generated fake images. GenImage has advantages over the existing fake image detection datasets in terms of the number of images, image content, and generator selection. With GenImage, the researchers can train detectors and compare their performance.

2. Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
The Huawei Noah's Ark Lab created the dataset.

3. What support was needed to make this dataset? who funded the creation of the dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.
Huawei Noah's Ark Lab.

4. Any other comments?
No.

B Composition

5. What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
The images represent the object in ImageNet dataset, such as goldfish, and tiger shark.

6. How many instances are there in total (of each type, if appropriate)?
The GenImage contains 2,681,167 images, including 1,331,167 real and 1,350,000 fake images.

7. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

GenImage contains fake images from multiple generative models. The generative models can produce additional images for future GenImage version.

37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks.

8. What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
Each instance is a real image or AI-generated fake image.

9. Is there a label or target associated with each instance?
Yes. Each instance is associated with a label, which represents it is a real image or a fake image.

10. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.
No.

11. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.
No.

12. Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.
Yes. The data splits are described in the main paper. The data splits are based on ImageNet.

13. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
No.

14. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?
ImageNet dataset is a part of GenImage.

15. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.
No.

16. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
Yes. The GenImage contains ImageNet. Some content in ImageNet is considered to be offensive. See: (1) https://excavating.ai/ (2) https://journals.sagepub.com/doi/full/10.1177/20539517211035955 (3) https://arxiv.org/abs/2006.16923

17. Does the dataset relate to people? If not, you may skip the remaining questions in this section.
Yes. There are some images containing people in ImageNet.

18. Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
No.

19. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
Yes. ImageNet contains natural persons.

20. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please describe how.
The GenImage contains ImageNet. Some content in ImageNet is sensitive in some way. See: (1) https://excavating.ai/ (2) https://journals.sagepub.com/doi/full/10.1177/20539517211035955 (3) https://arxiv.org/abs/2006.16923

21. Any other comments?
No.

C Collection Process

22. How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
The fake images are generated by the lables of ImageNet. The real images comes from ImageNet.

23. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created. Finally, list when the dataset was first published.
We estimate a timeline of 2 months for the data generation. The dataset is published in June, 2023.

24. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?
Midjourney API and Wukong API are used. Source codes of Stable Diffusion V1.4, Stable Diffusion V1.5, GLIDE, VQDM, BigGAN, ADM are used.

25. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
The dataset is not a sample from a larger set.

27. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?
The image is autonomous to generate.

28. Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
No.

29. Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.
Some people are in the ImageNet dataset.

30. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
We collect the data from ImageNet dataset.

31. Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
The fake data is not collected from individuals. The real data comes from the ImageNet dataset.

32. Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
The fake data is not collected from individuals. The real data comes from the ImageNet dataset.

33. If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate)
The fake data is generated by generative models and do not include personal data. The real data comes from the ImageNet dataset.

34. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
The fake data is not collected from individuals. The real data comes from the ImageNet dataset.

35. Any other comments?
No.

D Preprocessing/cleaning/labeling

36. Was any preprocessing/cleaning/labeling of the data done(e.g.,discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.
The preprocessing/cleaning/labeling of the data are mentioned in the paper.

37. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
No.

38. Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
Yes. The generators are provided in the GitHub page.

39. Any other comments?
No.

E Uses

40. Has the dataset been used for any tasks already? If so, please provide a description.
No.

41. Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
No.

42. What (other) tasks could the dataset be used for?
The dataset can be used for object classification and object detection.

43. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?
No.

44. Are there tasks for which the dataset should not be used? If so, please provide a description.
No.

45. Any other comments?
No.

F Distribution

46. Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
Yes. The dataset is publicly available.

47. How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
The GitHub address is https://github.com/GenImage-Dataset/GenImage.

48. When will the dataset be distributed?
The dataset has been distributed.

49. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
Yes. CC BY-NC-SA 4.0. https://github.com/GenImage-Dataset/GenImage/blob/main/License.

50. Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
No.

51. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
No.

52. Any other comments?
No.

G Maintenance

53. Who is supporting/hosting/maintaining the dataset?
Huawei Noah's Ark Lab.

54. How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
{zhumingjian, yunhe.wang}@huawei.com

55. Is there an erratum? If so, please provide a link or other access point.
No.

56. Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?
The first version of GenImage will not be updataed recently. If there exist any updates in the future, we will release a new sub-dataset as an complement of GenImage.

57. If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

The fake data is generated by generative models and do not include personal data. The real data comes from the ImageNet dataset.

58. Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.
Yes.

59. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.
We are open for discussion through the official contact methods.

60. Any other comments?
No.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[2] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[3] Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8060–8069, 2020.

[4] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII*, pages 86–103. Springer, 2020.

[5] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[6] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020.

[7] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

[8] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)*, pages 1–6. IEEE, 2019.