GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image Supplementary Material

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 Training and Testing on the Entire GenImage

To comprehensively assess the efficacy of GenImage, we train multiple models on the entire training set and evaluate these models on the entire testing set. We use 2,581,167 images for training and 100,000 images for evaluation. In Table 1, we demonstrate the binary classification results of all the methods. The results show that the models can achieve high testing accuracy when meeting the same generators in training.

Table 1: Results of different methods trained and evaluated on the entire dataset.

Method	ResNet-50	DeiT-S	Swin-T	CNNSpot	Spec	F3Net	GramNet
Accuracy(%)	99.5	99.8	99.9	98.3	98.0	92.1	99.2

B Training Details

For the different detectors, we use different training settings. These settings basically follow the official settings provided by these detectors.

ResNet-50 [2] [8] We use the training script provided in Timm Libiray [19]. The script in Timm is used for training ImageNet 1000 class images. We use a cosine annealing learning schedule to train a binary classification model. The number of epochs is 200, and the learning rate is 0.05. The batch size is 64. SGD optimizer is utilized. We enable Jensen-Shannon Divergence and CE loss for training. We use random erasing data augmentation, and the random erase probability is 0.6.

DeiT-S [3] [16] For training the detector, we utilize AdamW as the optimizer. We train the model for 300 epochs, and the batch size is 256. The learning rate for training is 0.0002. The warm-up epochs are 5 and warm-up learning rate is 1×10^{-6}. The random erase probability is 0.25.

Swin-T [4] [11] AdamW is also used for training Swin-T. The training epochs are 300. The batch size is 128. The learning rate is 0.0005. The warm-up epochs are 20, and the warm-up learning rate is 5×10^{-7}. The random erase probability is also 0.25.

CNNSpot [5] [18] We use Adam to train CNNSpot. The epochs for training are 30. The batch size is 64, and the learning rate is 0.0005. The JPEG probability and the blurring probability are 50%.

*Corresponding Author

[2] https://github.com/huggingface/pytorch-image-models/tree/v0.6.12/timm

[3] https://github.com/facebookresearch/deit

[4] https://github.com/microsoft/Swin-Transformer

[5] https://github.com/PeterWang512/CNNDetection

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

Spec [6] [21] We use FFT feature and SGD optimizer to train the model. The number of epochs is 64. The learning rate is 0.0002. The learning rate decay ratio is 10^{-2}. The weight decay value is 0.0001.

F3Net [7] [14] The training epochs are 5. The batch size is 12. The optimizer is Adam, with a learning rate of 0.0002. We use a binary cross entropy loss that comes inside a sigmoid function.

GramNet [8] [12] We train the model for 5 epochs, and the batch size is 14. The learning rate is 0.00001. Adam is used as the optimizer. The value of weight decay is 10^{-4}. The negative log likelihood loss is used for training a binary classification model.

C Analysis of Robustness against Adversaries

We perform new experiments to analyze the robustness against adversaries in GenImage. We use Fast Gradient Sign Method (FGSM) [7] and Wagner L2 Norm Attack(CW-L2) [2] for attacking methods. The results are shown in the Table 2. This analysis is performed on Stable Diffusion V1.4. We can see that the ResNet-50 is more robust for CW-L2 than FGSM.

Table 2: Results of Evaluating Models on Different Adversaries.

	FGSM	CW-L2
ResNet50	88.0	99.9

D Evaluation on Other Kinds of Generators

We demonstrate the ResNet-50 that are trained on all generators and evaluated on NVAE [17], CogView2 [6], StyleGAN [10], and IF [1], as shown in Table 3. For each generator, we collect 1000 real images and generate 1000 fake images. For CogView and IF, we use the images from ImageNet and the input sentences follow the template "photo of class", with "class" being substituted by ImageNet labels. We use a NVAE model pretrained on FFHQ [10] to generate fake images, and the real images come from FFHQ. StyleGAN is pretrained on LSUN bedroom, and the real images also come from LSUN [20]. The results show that the detector trained on our dataset can generalize to other kinds of generators.

Table 3: Results of Evaluating Models on Other Kinds of Generators.

	NVAE	CogView2	StyleGAN	IF
ResNet-50	93.4	97.5	97.9	90.2

E Evaluation on Different Number of Generators

We compare the ResNet-50 model trained on three different settings, as shown in Table 4. The generality of the detector increase as the number of generators increases. One generator is SD V1.4. Four generators are SD V1.4, Midjourney, BigGAN, and ADM. Eight Generators are all the generators in GenImage.

Table 4: Results of Evaluating Model on Different Number of Generators.

	NVAE	CogView2	StyleGAN	IF
One Generator	64.2	79.0	65.7	62.4
Four Generators	70.4	95.6	68.1	82.8
Eight Generators	93.4	97.5	97.9	90.2

[6] https://github.com/ColumbiaDVMM/AutoGAN

[7] https://github.com/yyk-wew/F3Net

[8] https://github.com/liuzhengzhe/Global_Texture_Enhancement_for_Fake_Face_Detection_in_the-Wild

F Evaluation in More Complex Scenes

The generalization performance of GenImage has been evaluated on Face and Art Images. We further conduct experiments on images generated by richer text descriptions and image-to-image generation, as shown in the Table 5. For obtaining richer text descriptions, we collect 1000 prompts and 1000 images from CC12M [3]. For image to image generation, we input the real images from ImageNet and a template of "a painting of class" to the generator. The generator used in the above experiments is Stable Diffusion. For obtaining multi-object content images, we use a template of "photo of class A, class B, and class C". The classes and real images come from ImageNet. It can be seen that the models trained on our eight generators are sufficient to generalize well on more kinds of images.

Table 5: Results of Evaluating Model in More Complex Scenes.

	Richer Text Descriptions	Image-to-Image Generation	Multi-Object Content
ResNet50	99.9	99.3	99.9

G Details of Generators

The guidance scale of SD1.4, SD1.5, GLIDE, VQDM, and ADM are 7.5,7.5, 3.0, 1.0, and 10.0. The sampling method of SD1.4, SD1.5, and ADM are DDIM. The steps of diffusion of SD1.4, SD1.5, ADM, GLIDE, VQDM are 50,50, 1000,127,100. The random seed of SD1.4 and SD1.5 is 42. Some hyperparameters of these generators are not specified. We use the API of Midjourney and Wukong. Thus we cannot know the details of their hyperparameters.

H Limitations

In terms of limitation, the GenImage dataset uses the same classes as ImageNet. As the real world contains an increasing variety of objects, although GenImage already covers plenty of object classes, our image data certainly does not cover the new objects emerging in the future. However, fine-tuning the pre-trained model on the data of these objects can solve this problem. We have shown the detector trained on GenImage can perform well on the generated face images. We further run test for demographic subgroups, as shown in Table 6. Each subgroup contains 1000 real images and 1000 fake images generated by Stable Diffusion V1.4. For images of children, women, racialized persons, and elderly people, we collect real images from Fairface [9]. Race group contains White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino. We collect real images of those living with disabilities from BGVP [15]. ResNet-50 trained on GenImage can perform well on these images. We have conducted experiments to explore the problem of biased model as much as possible. However, GenImage is not a dataset specially designed for face forgery and we do not focus on solving the problem of face image distribution. To this end, the detector trained on GenImage still has the potential to be biased for particular demographics.

Table 6: Results of Evaluating Models on Different Demographics.

	Children	Women	Racialized Persons	Elderly People	Those Living with Disabilities
ResNet-50	99.9	99.8	99.9	99.9	99.8

I Societal Impacts

There are many practical applications for the GenImage dataset, which will lead to significant social impact. When a detector is trained on it, the user has the ability to identify AI-generated content. The user runs the detector to obtain the prediction, and this action leads to consequences. For example, a teacher might run their students' submitted homework through a detector to determine if the student has cheated by using an AI image generator to complete their assignment. If this dataset were to be widely disseminated and used to train detectors, it could be incorporated into software designed for this task. In this situation, the accuracy and reliability of the detectors will largely rely on the

GenImage. Besides, the models trained on this dataset may show some undesirable tendencies on some new objects in the world. Although in our analysis, the models trained on our dataset have the ability to generalize, special attention is still recommended for practical use.

J Ethics Statement

In terms of ethics, "Datasheets for Datasets" has been uploaded to the supplementary materials. Our dataset is based on ImageNet. No additional personally identifiable information or sensitive personally identifiable information is introduced during the production of fake images in the GenImage dataset. During the dataset production, we do not introduce extra information containing exacerbated bias against people of a certain gender, race, sexuality, or who have other protected characteristics. The ethical issues in the ImageNet dataset have been discussed in previous works. Crawford et al. [4] explore the problems in ImageNet. The first problem is that all taxonomies or classificatory systems are political. For example, only "male" and "female" bodies are "natural." "Hermaphrodite" is offensively situated within the branch Person > Sensualist > Bisexual > alongside the categories "Pseudohermaphrodite" and "Switch Hitter." The second problem is the images of real people are often offensive. The third problem is that ImageNet's creators use people's photos without their knowledge. Denton et al. [5] find that assumptions around ImageNet generally rely on three themes: the aggregation and accumulation of more data, the computational construction of meaning, and the rendering of certain types of data labor invisible. There exists a dual ideological formation in these discourses: first around the accumulation of data and second around the disembodied, decontextualized nature of annotation work. Prabhu et al. [13] survey the threats of ImageNet. First, the reverse image search engines enable uncovering the "real-world" identities of the humans of the ImageNet dataset, which make them lose their privacy. Second, ImageNet paves the way for the emergence of even larger and more opaque datasets. Third, A Creative Commons license only tackles copyright issues - not privacy rights or consent to use images for training. ImageNet, which has been built on top of the Creative Commons, interprets it as a free for all, consent-included green flag.

K Hosting and Maintenance Plan

The authors will ensure the long-term maintenance of the GenImage dataset. The codes utilized in this research are based on third-party open-source codes. Therefore, we only provide open-source URLs, and we do not maintain these codes in this work. The dataset website is hosted on Github Pages [9]. GitHub is a prominent website hosting service. We provide comprehensive information about GenImage, including dataset introductions, dataset links, sample images, key performance evaluations, and terms of use. All these resources are accessible on an open platform and freely available for download by the public. The storage of the datasets will be facilitated through Baidu Cloud, a widely utilized cloud storage service in China.

L License

Our released datasets are under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International Public License ("CC BY-NC-SA 4.0"). The information on CC BY-NC-SA 4.0 can be accessed at the website [10]. The user who uses the datasets is considered as agreeing to comply with CC BY-NC-SA 4.0 and the dataset terms. GenImage is used for non-commercial purposes only, such as academic research and scientific publications. We do not allow the user to use the dataset for commercial purposes, such as selling data for commercial profits.

M Visulization of Generated Images

From Figure 1 to Figure 8, we show some examples in our dataset. We demonstrate 180 images for each generator. It can be seen that each generator can generate images with high diversity.

[9] https://github.com/Andrew-Zhu/GenImage

[10] https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode

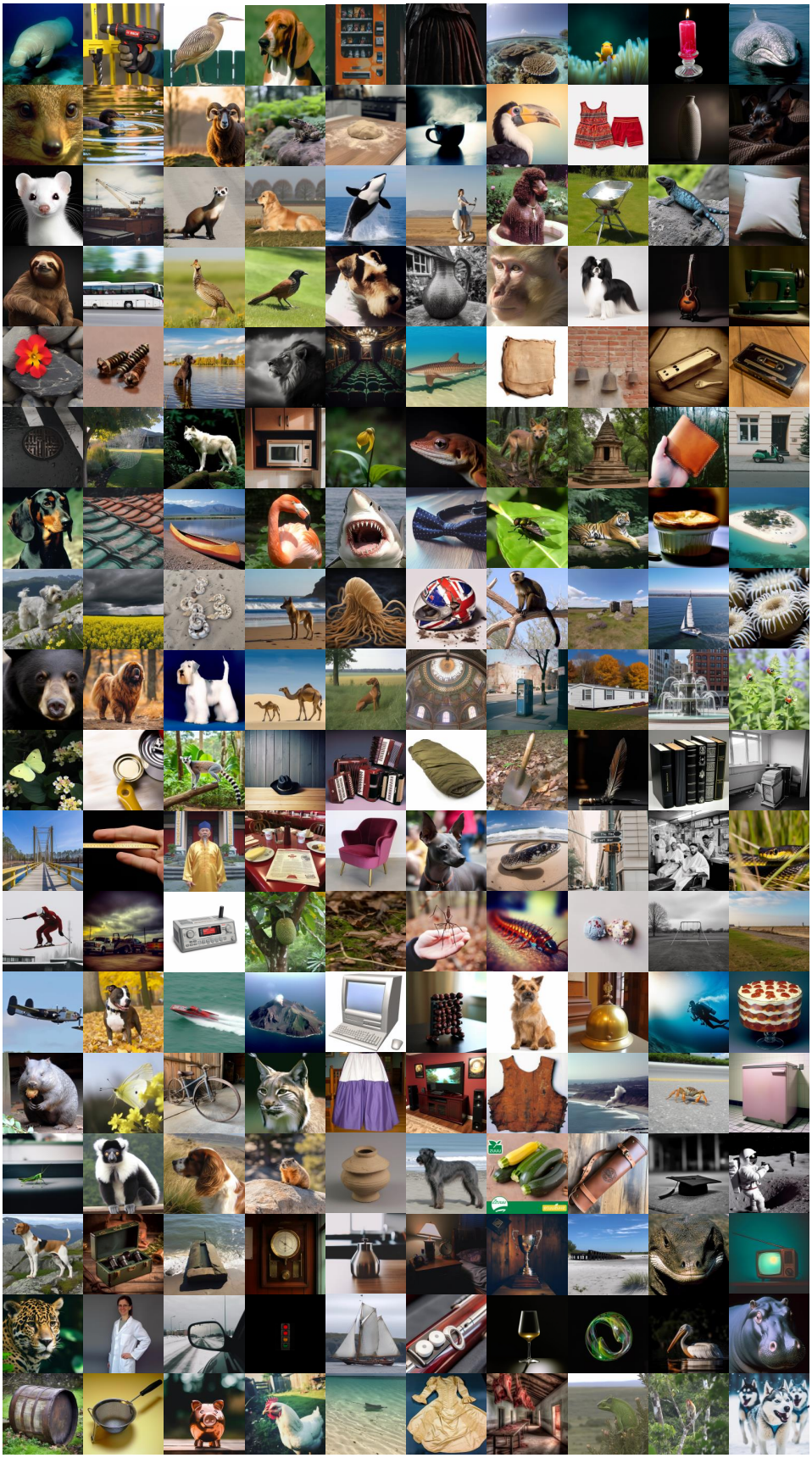

Figure 1: Midjourney Images

Figure 2: Stable Diffusion V1.4 Images

Figure 3: Stable Diffusion V1.5 Images

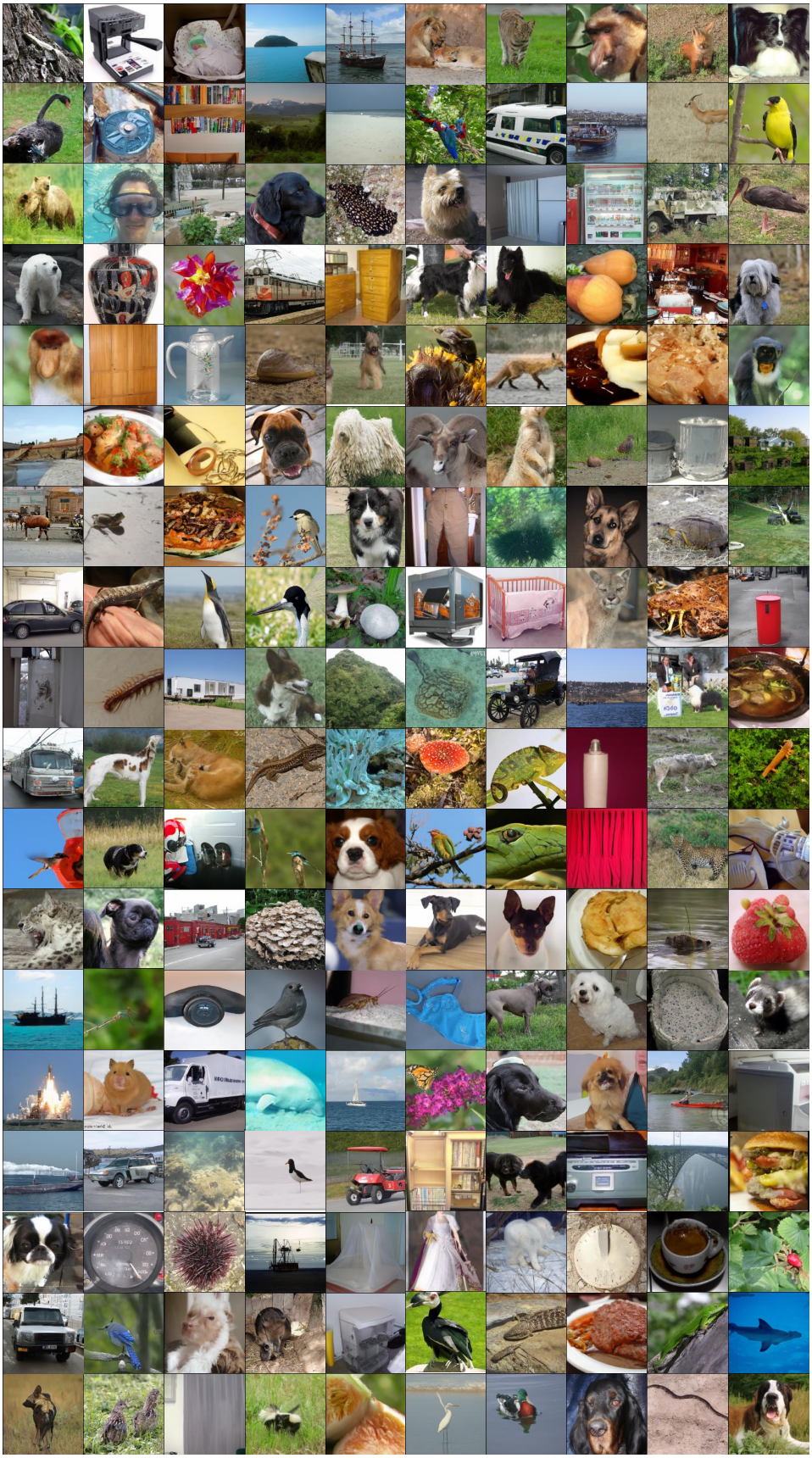

Figure 4: ADM Images

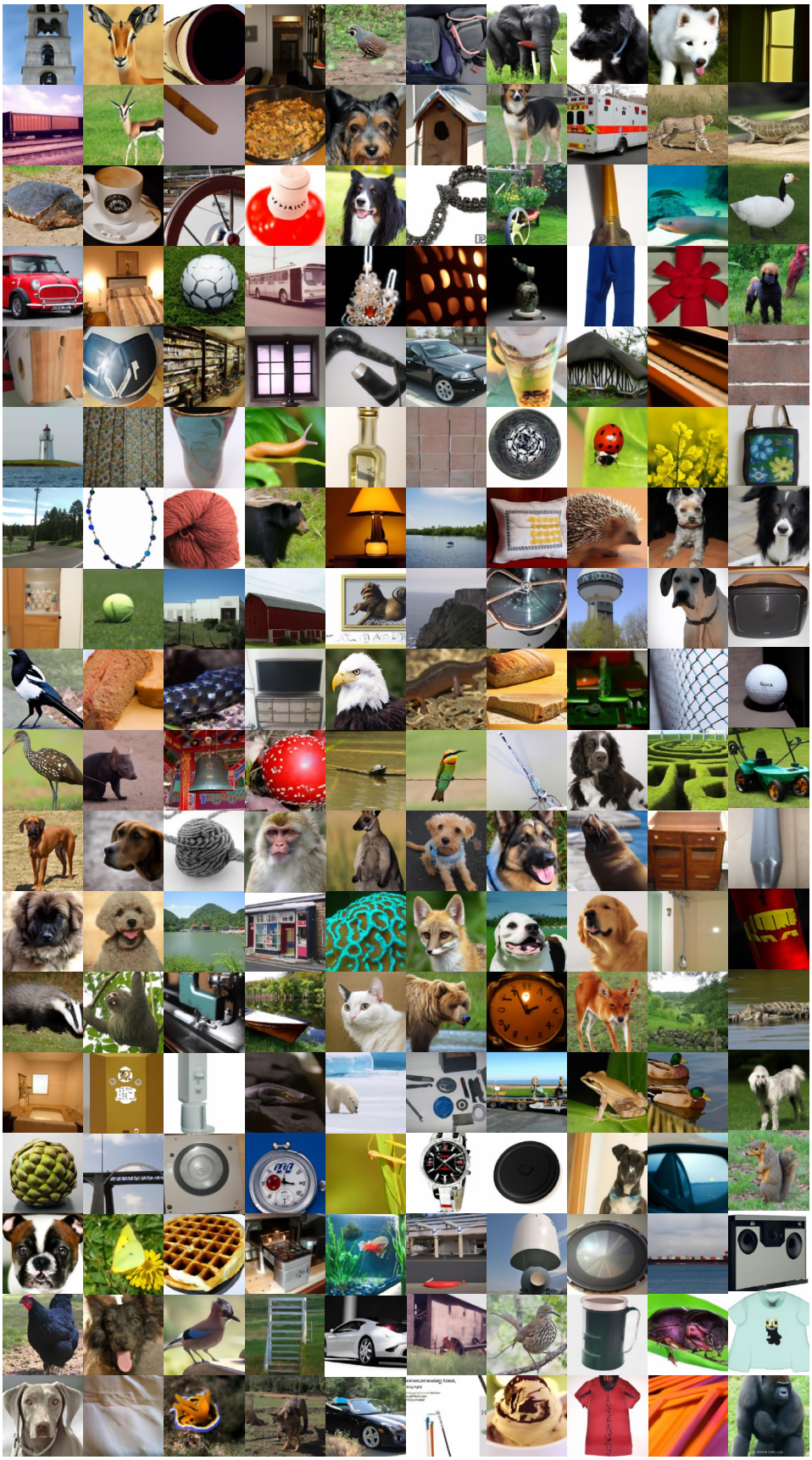

Figure 5: GLIDE Images

Figure 6: Wukong Images

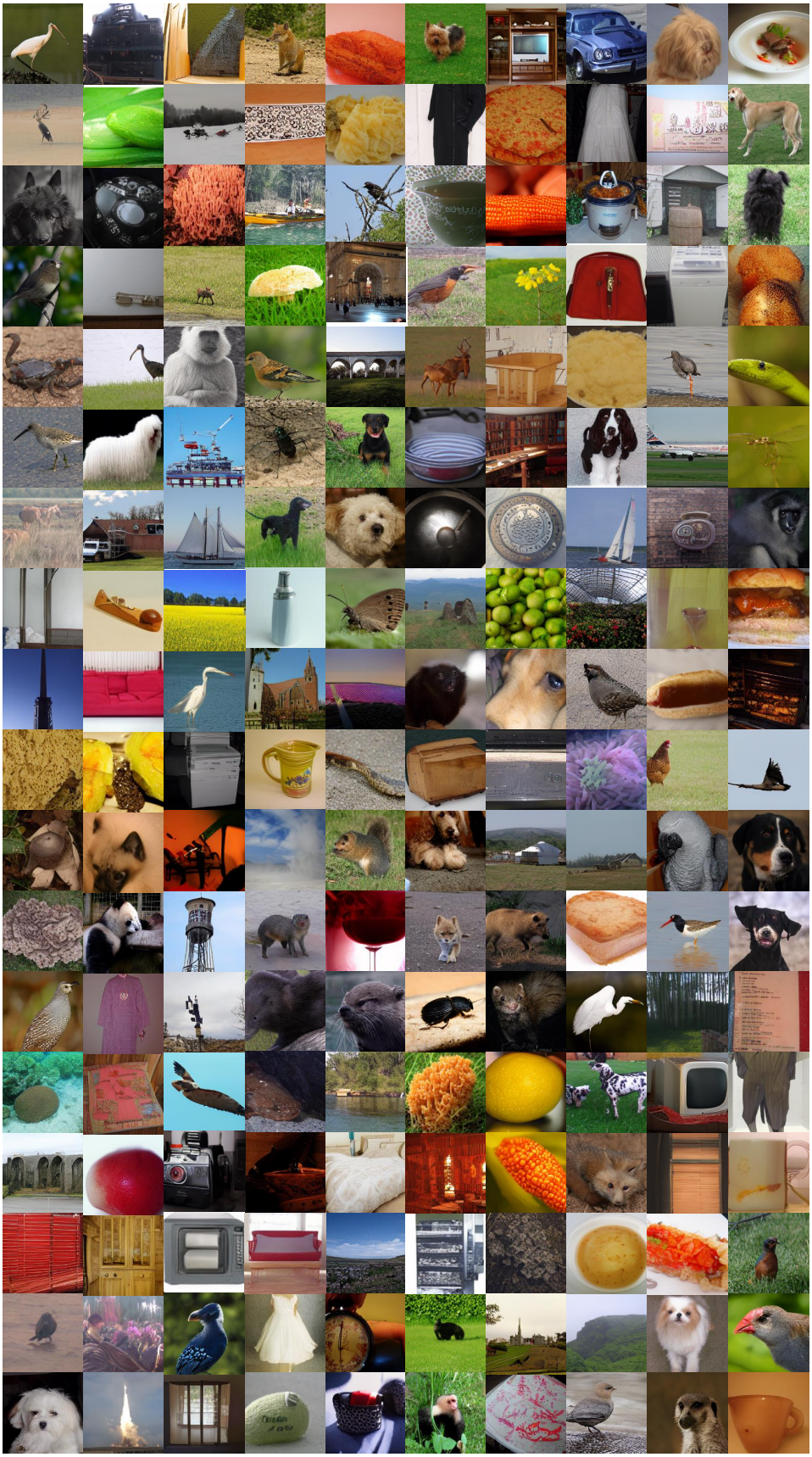

Figure 7: VQDM Images

Figure 8: BigGAN Images

References

[1] https://github.com/deep-floyd/if/tree/develop. 2023.

[2] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

[3] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.

[4] Kate Crawford and Trevor Paglen. Excavating ai: The politics of images in machine learning training sets. *Ai & Society*, 36(4):1105–1116, 2021.

[5] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, and Hilary Nicole. On the genealogy of machine learning datasets: A critical history of imagenet. *Big Data & Society*, 8(2):20539517211035955, 2021.

[6] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35:16890–16902, 2022.

[7] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[8] 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.

[9] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1548–1558, 2021.

[10] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[11] 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.

[12] 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.

[13] Vinay Uday Prabhu and Abeba Birhane. Large image datasets: A pyrrhic win for computer vision? *arXiv preprint arXiv:2006.16923*, 2020.

[14] 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.

[15] Devansh Sharma, Tihitina Hade, and Qing Tian. Comparison of deep object detectors on a new vulnerable pedestrian dataset. *arXiv preprint arXiv:2212.06218*, 2022.

[16] 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.

[17] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679, 2020.

[18] 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.

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

[20] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[21] 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.