Coursework Question 2: Creating a datasheet

3.2 Composition

- **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 be- tween them; nodes and edges)? Please provide a description.

All instances in the FFHQ dataset are PNG images of human faces of resolution 1024 x 1024 pixels and contain considerable variation in terms of age, ethnicity and image background and cover a good range of accessories like eyeglasses, sunglasses, hats etc.

- **How many instances are there in total (of each type, if appropriate)?**

There are 70,000 instances of the PNG images in total.

- **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).

The dataset is a sample of instances of human faces from the larger set of all images on Flickr. Various automatic filters were used to prune the set and Mechanical Turk was used to remove any images that were not human faces such as statues, paintings, or photos of photos.

- **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 PNG image of resolution 1024 x 1024 pixels. The images were crawled from Flickr (thus inheriting all the biases of that website) and automatically aligned and cropped. Only images under permissive licenses were collected. Various automatic filters were used to prune the set and finally Mechanical Turk was used to remove the occasional statues, paintings, or photos of photos.

- **Is there a label or target associated with each instance?** If so, please provide a description.

There are no labels associated with the instances.

- **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.

In relation to the distribution of training data, areas of low density are poorly represented making it more difficult for the generator to learn. This is an ongoing challenge in all generative modelling techniques however others have shown that drawing latent vectors from a truncated (A. Brock, 2019), (Marchesi, 2017) or otherwise shrunk (Dhariwal, 2018) sampling space tends to improve average image quality, although some variation is lost. Avoiding sampling from the extreme regions of the sample space using the truncation trick (see Section 7 in (Tero Karras, 2021)) can help tackle this problem and for the FFHQ dataset this leads to a sort of an average face.

For use cases that require separate training and validation sets, the creators appointed the first 60,000 images to be used for training and the remaining 10,000 for validation.

- **Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.

All images were crawled from Flickr, hence inherit all the biases of that website.

- **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.

The instances represent a diverse range of human faces of different ages, gender, ethnicity etc but do not contain labels identifying specific instances with any subpopulations.

- **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, all images are of real people from the Flickr website. Only images under permissive licenses were collected.

- **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race 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 provide a description.

No.

3.3 Collection Process

- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?** How were these mechanisms or procedures validated?

A tool was used to scrape the images from the Flickr website.

- **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 FFHQ data was collected by researchers Tero Karras, Samuli Laine and Timo Aila as part of this research paper (Tero Karras, 2021).

- **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.

Unknown to the author of the datasheet.

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

The data was collected via a third party website, Flickr.

- **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 individual images were published in Flickr by their respective authors under either Creative Commons BY 2.0, Creative Commons BY-NC 2.0, Public Domain Mark 1.0, Public Domain CC0 1.0, or U.S. Government Works license. All of these licenses allow free use, redistribution, and adaptation for non-commercial purposes.

- **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).

Yes. On the GitHub repository there are instructions in the README.md file on how to find out whether your photo is included in the FFHQ dataset using this link to search with your Flickr username, and how to get your photo removed.

3.5 Uses

- **Has the dataset been used for any tasks already?** If so, please provide a description.

Yes. The original paper used the FFHQ dataset in their work on style-based GAN generator architectures (Tero Karras, 2021). The paper also has 22 citations on IEEE Xplore.

- **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.

Yes. The data is publicly available on GitHub at https://github.com/NVlabs/ffhq-dataset.

- **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 dataset consumer 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 risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

The creators of the dataset ensured that there is minimal risk. See answer on consent in Section 3.3. Some of the permissive licenses require giving appropriate credit to the original author, as well as indicating any changes that were made to the images. The license and original author of each image are indicated in the metadata.

- **Are there tasks for which the dataset should not be used?** If so, please provide a description.

This dataset is not intended for, and should not be used for, development or improvement of facial recognition technologies. For business enquiries, please visit NVIDIA Research Licensing.

Works Cited

A. Brock, J. D. a. K. S., 2019. *Large scale GAN training for high fidelity natural image synthesis,* s.l.: Proc. Int. Conf. Learn. Representations.

Dhariwal, D. P. K. a. P., 2018. *Glow: Generative flow with invertible 1x1 convolutions,* s.l.: Proc. Int. Conf. Neural Inf. Process. Syst..

Marchesi, M., 2017. *Megapixel size image creation using generative adversarial networks,* s.l.: CoRR, vol. abs/1706.00082.

Tero Karras, S. L. ,. a. T. A., 2021. *A Style-Based Generator Architecture for Generative Adversarial Networks,* s.l.: IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 43, NO. 12.