Template for Coursework Question 2: Creating a datasheet

3.2 Composition

Dataset creators should read through these questions prior to any data collection and then provide answers once data collection is complete. Most of the questions in this section are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for their chosen tasks. Some of the questions are designed to elicit information about compliance with the EU's General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions.

Questions that apply only to datasets that relate to people are grouped together at the end of the section. We recommend taking a broad interpretation of whether a dataset relates to people. For example, any dataset containing text that was written by people relates to people.

- **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.
    - The main portion of the dataset consists of human faces with variations in people's age, ethnicity, and image background. Also included are a series of 68 facial landmarks that have been identified with the use of the dlib library package.
- **How many instances are there in total (of each type, if appropriate)?**
    - The original and only version (2019) of the dataset includes a total of three sets of the same 70,000 images, with varying image sizes, and a series of features/landmarks included for each image.
- **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 comprises a sample of varying instances for faces, described as having good coverage for age, ethnicity, and accessories (hats, glasses, etc). As the images are collected from Flickr, it is recognised that the dataset is subject to any bias of the website. Details for verification of coverage have not been provided.
- **What data does each instance consist of?** "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
    - The dataset consists of three sets of the same RGB PGN images with varying dimensions. The first set contains the pre-processed images at their native resolution. The dlib library package has been applied to the second set to set a resolution of 1024x1024 throughout. The final set contains thumbnail images of size 128x128. The 68 face landmarks are included in the form of a string that returns from the dlib library.
- **Is there a label or target associated with each instance?** If so, please provide a description.
    - The facial landmarks provide target data in the form of locations that can be used as ground truths for model training.

- **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.
    - A split of 60,000 (85.7%) images for training and 10,000 for validation is detailed for use cases of separate data channels. Specific reasons for these values are not provided.
- **Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.
    - The set of raw uncropped images is recognised as containing multiple examples of the same instance, although they have been removed from the main dataset.

If the dataset does not relate to people, you may skip the remaining questions in this section.

- **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.
    - Although subpopulations are included in the dataset, there are no explicit references to this type of data.
- **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.
    - Only images have explicitly been collected for this dataset. It may, however, be possible to infer information such as location depending on image backgrounds.
- **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

As with the questions in the previous section, dataset creators should read through these questions prior to any data collection to flag potential issues and then provide answers once collection is complete. In addition to the goals outlined in the previous section, the questions in this section are designed to elicit information that may help researchers and practitioners to create alternative datasets with similar characteristics. Again, questions that apply only to datasets that relate to people are grouped together at the end of the section.

- **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?
    - The full set of images has been collected by the NVIDIA team from the online image host Flickr. Specific information on any software used for image acquisition has not been provided.
- **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)?**
    - Data collection processes were created and run by researchers from NVIDIA. Data labelling was completed with the use of automated software and therefore did not involve crowdworkers.

- **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.
    - There are currently no details of any ethical reviews having been completed for the dataset or the collection methods applied.

If the dataset does not relate to people, you may skip the remaining questions in this section.

- **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**
    - As described previously, the data was collected indirectly via a 3rd party image hosting service.
- **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.
    - No. However, the data collected is, to the best knowledge of the authors, intended for free use and has exclusively been made available by the author of the content.
- **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).
    - As described in the previous question, consent has not been collected. However, the authors provide means for data to be removed from the dataset with proof of ownership through the link: https://nvlabs.github.io/ffhq-dataset/search/

3.5 Uses

The questions in this section are intended to encourage dataset creators to reflect on the tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers to make informed decisions, thereby avoiding potential risks or harms.

- **Has the dataset been used for any tasks already?** If so, please provide a description.
    - Currently, only the original paper could be identified (Karras, Laine and Aila, 2018), in which an alternative architecture for generative adversarial networks is proposed. In this research, the dataset is used for the separation of attributes such as pose and identity.
- **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.
    - A collection of papers that use the dataset are available through Cornell University (https://arxiv.org/)
- **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?

    - All the data collected was from the public domain, therefore carrying a low risk for harm. However, there may still be a risk of personal information in some of the images.
- **Are there tasks for which the dataset should not be used?** If so, please provide a description.
    - The dataset is detailed as not for use for the creation or improvement of any facial recognition systems.

References

Karras, T., Laine, S. and Aila, T., 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4401-4410).