**Coursework Question 2: Creating a datasheet**

**--- 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.

These instances are unique photographs of people’s headshots. It consists of people that have variation between many personal factors such as age and ethnicity, as well as photograph variations such as image background. There is also a range of accessories that are included in the photos such as sunglasses, eyeglasses, and hoods among others. We also have no interactions between photographs, in terms of composition, throughout the dataset.

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

This dataset consists of 70,000 photo instances at a 1024x1024 resolution with no information on the split of personal features such as age, gender, and ethnicity.

**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).

This dataset is not a sample of instances of a large set, it contains all possible instances that were legally allowed to be used freely. However, there are versions of the FFHQ dataset: one that contains instances for the processed images and one that contains instances that haven’t been cropped or aligned. Both sets of the data contain the same number of instances, and the intention is that these datasets are used for different types of problems.

**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 photographs are raw images crawled from Flickr that have been automatically aligned and cropped using dlib, and Amazon Mechanical Turk was used to remove photos of photos, status or paintings – cases where there was no ‘person’ in the photograph. The package dlib uses the eyes and nose as a three-point system to align the face. However, this is the only pre-processing that has been done.

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

There are no labels or targets associated with each observation.

**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.

The data is split into ‘training’ and ‘validation’ datasets which is given to us in the ‘category’ key in the .json file for each instance. The training dataset makes up 85.7% of the dataset with the validation dataset containing the remaining 14.3%. Since the instances are not ordered, the split was done by taking the first 60,000 as the training set and the remaining 10,000 as the validation set.

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

The only candidate for noise in the dataset is the non-uniform backgrounds for each image. Since the profile shots are in a range of locations of varying contrast to the individual’s

**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.

There are no specific identifying features of any subpopulations in this dataset; however, it is possible to identify subpopulations visually, such as gender and age group, for each instance.

**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.

This dataset does contain two features that could be used to identify the subject: author and photo\_url. These two features would have varying impact depending on the instance itself. In the case of author, this could be the subject’s name for that particular instance which is a direct identification feature itself. Regarding the photo\_url, this is the Flickr URL that the photo has been crawled from which could be indirectly used as an identification feature since that profile could contain personal, identifying information.

**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.

The only direct, sensitive feature in this dataset is ‘country’ which describes where the photograph was taken which is sensitive from a location point of view. There are no other direct features in this dataset regarding sensitive features such as but not limited to race, ethnicity, and crime history.

**--- 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?

Each instance in the dataset was crawled from the photograph website, Flickr, using an automated system.

**Who was involved in the data collection process (e.g., students, crowd workers, contractors) and how were they compensated (e.g., how much were crowd workers paid)?**

This data collection was done internally by NVIDIA by employees of the company, so they were compensated as per their employee contracts. Jaakko Lehtinen, David Luebke, and Tuomas Kynkäänniem formed a feedback team; Janne Hellsten, Tero Kuosmanen, and Pekka Jänis created the compute infrastructure; and Vahid Kazemi and Josephine Sullivan researched the pre-processing facial alignment techniques.

**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 was no ethical review process conducted by an intuitional review board.

**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 indirectly through Flickr. The individual would have to had uploaded the photograph to Flickr, allow it to be under a permissive licence, have made the photograph public and have not tagged it to excluded from computer vision projects.

**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 individuals in questions did consent to the use of their data. Images were only used if they were under a permissive licence and intended for free use.

**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).

Individuals have the option to remove their photograph from the dataset by doing one of the following on Flickr: mark the photograph as not to be used for computer vision problems, change the licence of the photograph, set the photo as private, or remove the photograph from Flickr completely. Once this has been done, notifying NVIDIA will remove the individual’s photograph from the dataset.

**--- Uses ---**

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

The dataset has been used in 445 papers, typically, but not always, using Generative Adversarial Network models in the context of image quality improvement, image generation and image inpainting.

**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.

The links to all the papers and benchmarks that have been defined from the dataset can be found at the following URL: https://paperswithcode.com/dataset/ffhq.

**Is there anything about the composition of the dataset or the way it was collected and pre-processed/cleaned/labelled 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?

This is nothing about the composition of the dataset or the way that it was collected or processed that might impact future uses.

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

This dataset should not be used for facial recognition tasks due to the permissions and licencing restrictions.