Dataset description

Data selection process :
The FFHQ dataset is downloaded folder by folder from the official depot on the authors' Google Drive :
https://drive.google.com/drive/folders/1tZUcXDBeOibC6jcMCtgRRz67pzrAHeHL?usp=sharing
The folders are unzipped and all images with licenses corresponding to :

- https://creativecommons.org/licenses/by-nc/2.0/ and
- http://www.usa.gov/copyright.shtml

are removed. For each existing combination of the Sk ; A ; Se ; C ; P attributes (as described below), we browse the resulting data sequentially (in the officially available order) and take the first picture corresponding to the required combination. This gives us a dataset of 72 RGB images of resolution 1024*1024. We then use the Lanczos resizing function of the PIL library (https://arxiv.org/abs/2104.11222) to resample the dataset to a dimension of 512*512, and rename all images using the process described below.

Data naming convention :
<Sk>_<A>_<Se>_<C>_<P>_<B>_<Hc>_<D>_<Hs>.png

Sk : **Skin tone index**

0 = pale (I/II on Fitzpatrick's scale)
1 = intermediate (III/IV on Fitzpatrick's scale)
2 = dark (V/VI on Fitzpatrick's scale)

Labeling and selection comments :

- equally distributed in the dataset
- ambiguities are generally avoided by avoiding to select skin tone-ambiguous pictures or pictures with extreme lighting conditions
- remaining ambiguities caused by lighting conditions are solved by the following priority order : dark > intermediate > pale

A : **Age index**

0 = younger (babies and children)
1 = intermediate (teenagers and young adults)
2 = older (older adults and elderly)

Labeling and selection comments :

- equally distributed in the dataset
- ambiguities are generally avoided by avoiding to select age-ambiguous pictures

Se : **Biological sex index**

0 = male

1 = female

Labeling and selection comments :

- equally distributed in the dataset
- biological sex of the person determined based on physical cues represented on the photo

C : **Corpulence index**

0 = less corpulent

1 = more corpulent

Labeling and selection comments :

- equally distributed in the dataset
- as corpulence standards are not comparable across people of different sex, populations or ages, we define the images as “chosen in such a way that, for all samples of a similar skin tone, sex and age, all those tagged as less corpulent are visibly less corpulent than all those tagged as more corpulent”. Care is taken that the difference between both groups is clearly visible in most cases.

P : **Particularity index**

0 = absence of particularity

1 = presence of particularity

Labeling and selection comments :

- equally distributed in the dataset
- particularities are defined as :
    - pieces of clothes or accessories ;
    - scars, stains or birthmarks ;
    - beards, mustaches or other features of facial hair ;
    - tattoos or visible makeup

  so long as those features occult or obscure part of the face of the main subject on the image.

B : **Bangs index**

0 = absence of bangs

1 = presence of bangs

Labeling and selection comments :

- bangs are defined as strands of hair covering more than 40% of the forehead’s height and more than 80% of the forehead’s width, falling relatively straightly or sometimes in a slightly curved manner and cut before they reach the eyes, potentially shorter than the rest of the hair

Hc : **Hair colour index**

0 = black

1 = blond

2 = brown

3 = gray

bald = bald

Labeling and selection comments :

- ambiguities caused by multiple colours are solved by the following priority order : black > brown > blond > gray. White hair is considered gray.
- baldness is defined by the point from which a change in hair colour and shape would no longer be obvious to the naked eye
- baldness according to this attribute is shared with baldness according to the hair shape (Hs) attribute

D : **Double-chin index**

0 = absence of double-chin

1 = presence of double-chin

Labeling and selection comments :

- a double-chin is defined as a lump of flesh going visibly below the chin, distinctly separated from a visible part of the chin and the neck by folds of the skin or shadowy areas (please note that on some images the remaining visible neck area is very small)

Hs : **Hair shape index**

0 = straight hair

1 = curly hair

bald = bald

Labeling and selection comments :

- curly hair include visibly wavy hair, with some hair have visible oscillations of two periods or more, as well as frizzy hair and hairstyles such as dreadlocks
- some pictures where very little hair is visible were labeled based on case-by-case decisions
- baldness is defined by the point from which a change in hair colour and shape would no longer be obvious to the naked eye
- baldness according to this attribute is shared with baldness according to the hair colour attribute

License and privacy :

As a derived work of FFHQ, this dataset falls under its license : Creative Commons BY-NC-SA 4.0. We additionally acknowledge their paper as the original piece of work which justified its creation : Karras et al., CVPR 2019, A Style-Based Generator Architecture for Generative Adversarial Networks (https://arxiv.org/abs/1812.04948).

The license of each individual picture can be retrieved using FFHQ’s publicly available metadata.

Just like the original authors, we are committed to protecting the privacy of individuals who do not wish their photos to be included. Please follow their explanations (https://github.com/NVlabs/ffhq-dataset#privacy) if you find that one of your photos is a part of this dataset against your will, and afterwards additionally contact us using our email address transferlearning-event@gmail.com so that we can remove the corresponding image from our data.