Improving Evaluation of Facial Attribute Prediction Models: Appendix

A Training Details

This section contains a complete list of the hyperparameters and training times for each model. To determine the fixed epoch counts, we trained until the validation loss stopped decreasing (rounded to a nearby multiple of 10 epochs). All models trained on the aligned version of CelebA use SGD with a batch size of 256, initial learning rate of 0.1, momentum of 0.9, weight decay of 0.0001, following the ResNet paper. Models are trained on a single NVIDIA GTX 1080 Ti. Our code was implemented using the Pytorch framework (detailed software requirements are provided with the code). Note that we use the official train/val/test splits provided with CelebA.

- ResNet-18: 40 epochs, learning rate multiplied by .9 each epoch.
- ResNet-18 (pretrained): 20 epochs, learning rate multiplied by .8 each epoch
- ResNet-18 (10%): 80 epochs, learning rate multiplied by .1 on validation loss plateau with a patience of 10
- ResNet-18 (10%, pretrained): 40 epochs, learning rate multiplied by .9 each epoch

For the unaligned data, we divide both batch size and initial learning rate by 4 to allow training on a single GPU. All models trained on the unaligned version of CelebA therefore use SGD with a batch size of 64, initial learning rate of 0.025, momentum of 0.9, weight decay of 0.0001.

- ResNet-18: 60 epochs, learning rate multiplied by .1 on validation loss plateau with a patience of 10
- ResNet-18 (pretrained): 20 epochs, learning rate multiplied by .9 each epoch
- ResNet-18 (10%): 80 epochs, learning rate multiplied by .1 on validation loss plateau with a patience of 10
- ResNet-18 (10%, pretrained): 40 epochs, learning rate multiplied by .9 each epoch

All other results discussed in the paper use the same training hyperparameters as the baseline models. For example, the pretrained ResNet-18 model using weighted BCE loss was trained using the same hyperparameters as the pretrained ResNet-18 model using unweighted BCE and similarly averaged over 5 runs.

Training times are provided below as the average number of seconds per epoch. Because pretraining does not impact train time, we provide times only for the four versions of the dataset. Note that these times include computation of validation results, so the 10% times aren't simply one tenth of the 100% times. Also note that multiple machines with mixed hardware were used for training. For simplicity, the times below are based on a single machine with an Intel Core i7-8700 CPU, 32 GB of RAM, and a Seagate ST1000DM010-2EP1 Hard Drive running Ubuntu 18.04.

- 100%, cropped and aligned: 551 ± 35s per epoch (6.1 hours for 40 epochs).
- 10%, cropped and aligned: 73 ± 2s per epoch (1.6 hours for 80 epochs).
- 100%, uncropped: 2279 ± 216s per epoch (38.0 hours for 60 epochs).
- 10%, uncropped: 335 ± 39s per epoch (7.4 hours for 80 epochs).

To improve ease of replicability, we provide the command line arguments we used to generate all results along with our code.

B Examples

To demonstrate some of the labeling issues present in CelebA, we provide three examples of poorly or inconsistently labeled attributes (*Bald*, *Receding Hairline*, and *Big Lips*). We also provide examples of the *wearing necklace* attribute to demonstrate the difficulties faced by the cropped and aligned images. All samples were generated by randomly sampling 20 images from the validation set labeled as that attribute, using a random seed of 0 to avoid sampling bias. Our code for sampling random validation images labeled with a particular attribute is provided alongside the training code.

Note that only half of the 20 images labeled as *bald* match the strictest definition (no hair on the scalp). In a larger sample of 200 images we only found 80 images meeting this definition. *Receding hairline* is more ambiguous, but many of the images contain hair that is clearly just tied or pulled back. It is possible that some labelers were interpreting "receding hairline" as "receding hair," e.g. the hair being pulled backward. However, even with this definition several are clearly mislabeled (several images contain close-cropped hair, which does not fit any definition of “receding hairline”). For the *big lips* attribute, it is generally unclear what distinguishes lip size in the sampled images from other images. For the *wearing necklace* attribute, a necklace is visible in only 7 of the sampled images. We checked the full-size versions of these images and found that there are 4 cases where the necklace was cropped out and 9 cases where no necklace is visible even in the full image (we do not have sufficient data to say how frequently *wearing necklace* is labeled incorrectly, but we note that this may explain why our pretrained model using the uncropped images only obtains an F1 of 59.03 for this attribute).

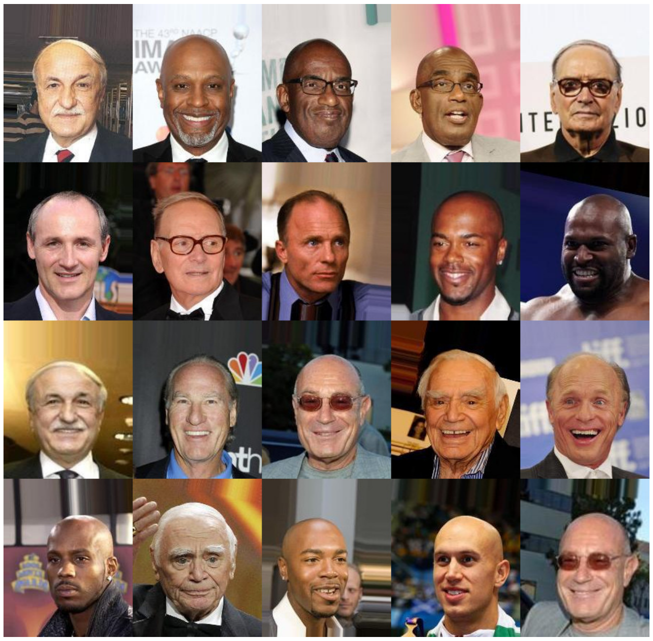

Figure 1: A random sample of 20 images in the validation set labeled with *Bald*.

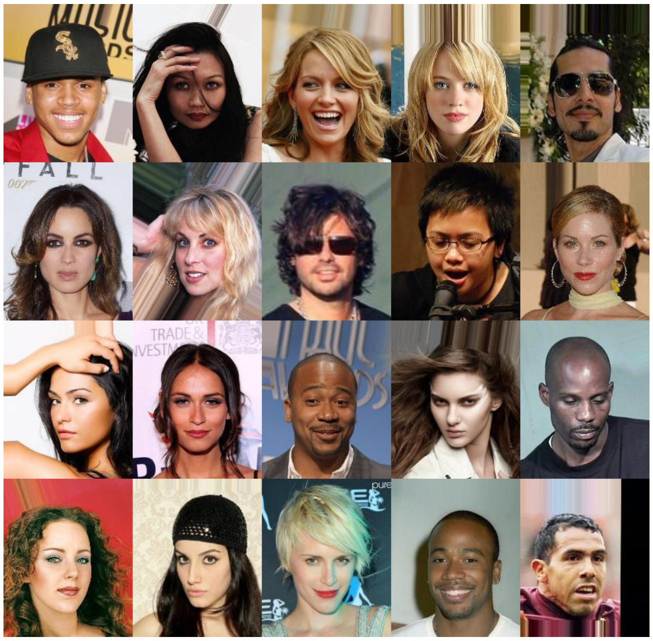

Figure 2: A random sample of 20 images in the validation set labeled with *Big Lips*.

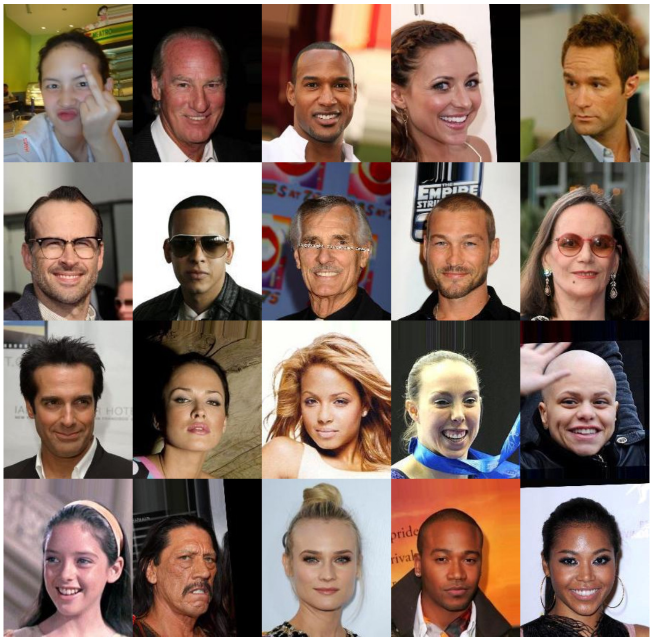

Figure 3: A random sample of 20 images in the validation set labeled with *Receding Hairline*.

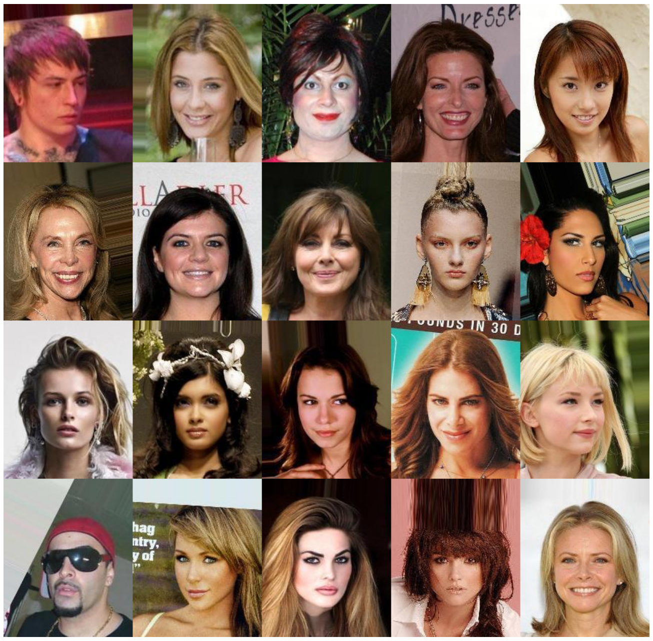

Figure 4: A random sample of 20 images in the validation set labeled with *Wearing Necklace*.