# Improving Evaluation of Facial Attribute Prediction Models

Bryson Lingenfelter and Emily M. Hand
University of Nevada, Reno

***Abstract*— CelebA is the most common and largest scale dataset used to evaluate methods for facial attribute prediction, an important benchmark in imbalanced classification and face analysis. However, we argue that the evaluation metrics and baseline models currently used to compare the performance of different methods are insufficient for determining which approaches are best at classifying highly imbalanced attributes. We are able to obtain results comparable to current state-of-the-art using a ResNet-18 model trained with binary cross-entropy, a substantially less sophisticated approach than related work. We also show that we can obtain near-state-of-the-art results on accuracy using a model trained with just 10% of CelebA, and on balanced accuracy simply by maximizing recall for imbalanced attributes at the expense of all other metrics. To deal with these issues, we suggest several improvements to model evaluation including better metrics, stronger baselines, and increased awareness of the limitations of the dataset.**

## I. INTRODUCTION

Facial attribute labels describe a face with natural language features such as *big nose*, *bushy eyebrows*, *gray hair*, and *smiling*. In addition to the direct utility of being able to describe a face in words, attribute labels have been used to improve face verification and identification [17][23], semantic segmentation [16], and other face parsing tasks such as detection and landmarking [20]. Facial attributes have also recently become popular for face editing [2][11]. The largest and most widely used facial attribute dataset is CelebA [19], which contains $202,599$ images of $10,177$ people labeled with 40 binary attributes. The images are provided in both the original, uncropped format and as $218 \times 178$ cropped and aligned images. We refer to the two versions as CelebA-ITW (In the wild) and CelebA-C+A (Cropped+aligned). Examples of both are shown in Figure 1.

In this work we show that near-state-of-the-art accuracy can be obtained on both versions of the CelebA dataset using a ResNet-18 model [10] trained with binary cross-entropy loss without any auxiliary data. This is in contrast to most recent attribute prediction approaches, which use substantially larger models and additional information such as segmentation masks and identity labels. By using initial weights pretrained on ImageNet, our results become even more competitive. On CelebA-ITW our results with pre-training substantially improve upon the accuracy obtained by current state-of-the-art models, most of which use auxiliary data far closer to the target domain.

We argue that a major reason models struggle to improve upon such a simple baseline is that the metrics used to evaluate them are severely flawed. Due to the imbalanced nature of the dataset, very high accuracy can be obtained for some attributes by a naive classifier which always predicts the majority class. We obtain results not far behind current state-of-the-art even when randomly discarding 90% of the training data, which disproportionately impacts the least balanced attributes. Furthermore, we show that balanced accuracy, used by several works as an alternative metric for dealing with these issues, can in fact be even worse for measuring performance on imbalanced data. We demonstrate how balanced accuracy can be exploited by training a model to a balanced accuracy score of 88.4%, only slightly behind state-of-the-art, with an average precision of just 58.6%. These metrics result in consistent overestimation of model quality, masking labeling issues which prevent reasonable performance on certain attributes. Better metrics show that several attributes are too subjective or poorly labeled to be reliably predicted.

Fig. 1. Examples of CelebA-ITW images (top) and their CelebA-C+A versions (bottom).

These flaws in currently used evaluation metrics, combined with the wide variety of backbone models and hyperparameter selections in other state of the art approaches as well as the lack of publicly available implementations, make it difficult to meaningfully compare different methods. To deal with this issue, we provide several suggestions for improved evaluation of facial attribute prediction models. Future work should evaluate models using F1-score or other metrics not affected by true negative counts, provide comparisons to stronger baselines more closely related to the proposed method, and better acknowledge the limitations of the dataset. Almost all labels are applied inconsistently across different images of the same person, suggesting that a new attribute dataset may be necessary for better evaluation of future work. We provide our implementation and per-attribute results at github.com/blingenf/celeba-baselines as a simple but strong baseline for future work to compare to.

This material is based upon work supported by the National Science Foundation under Grant No. 1909707.

978-1-6654-3176-7/21/$31.00 ©2021 IEEE

## II. RELATED WORK

Since the release of the CelebA dataset in 2015, there have been many proposed methods for CelebA attribute prediction. Liu et. al. used three deep Convolutional Neural Networks (CNNs) – $\text{LNet}_0$, $\text{LNet}_s$ and ANet – where the LNet networks detect the face in an unaligned image and ANet predicts attribute labels. Linear SVMs are then trained on the validation set to translate features learned by ANet to attribute predictions [19].

Later works rely on more typical end-to-end CNN models. MOON [21], which uses CelebA-C+A, consists of VGG-16 with a multitask loss function which accounts for differences between a source and target distribution. AFFACT [6], which provides results for both CelebA-C+A and CelebA-ITW (with faces detected by a pretrained face detector), uses ResNet-50 combined with both train-time and test-time augmentations. MCNN-AUX [8] uses a shallower CNN with different branches for different attribute groupings to take advantage of relationships between attributes.

Other works use additional data or labels to improve performance. SSP+SSG [15] takes advantage of the relationship between part localization and attribute prediction, using semantic segmentation to improve prediction performance. A semantic segmentation model trained on the segmentation-labeled Helen face dataset is used to gate and pool activations in a VGG-based architecture. Later work by the same authors uses an Inception-v3 backbone which jointly learns attribute prediction and semantic segmentation, improving the performance of both [16]. Segmentation data has also been used by [9], who use a Generative Adversarial Network (GAN) to generate segmentation masks which are then used to generate an additional set of features to combine with features from the RGB images.

In addition to auxiliary data, auxiliary labels can be used to improve attribute prediction. LMLE and CLMLE [12] deal with class imbalance by learning an embedding function which separates cluster distributions within and between classes. They use DeepID2 features trained on the CelebFaces+ dataset [22], which was used to create CelebA, effectively meaning that CelebA identity labels are auxiliary data. HFE [24] also takes advantage of the identity labels provided by CelebA by enforcing that representations should be separated by both attribute and identity information. Their method uses a DeepID2 backbone with fully-connected branches for each attribute. PS-MCNN [1] uses attribute groupings and with an additional shared network pretrained using identity labels. By combining attribute loss with identity loss (PS-MCNN-LC), they are able to obtain state-of-the-art results.

## III. BASELINE EXPERIMENTS

In this section we establish a simple baseline approach for facial attribute prediction. We then show that we are able to obtain results close to all state-of-the-art methods discussed in Section 2 following this approach, even when using far less data.

### A. Experimental Setup

For both CelebA-ITW and CelebA-C+A, we train one ResNet-18 model on the entire training set and another on a randomly sampled subset of 10% of the training set. We use the same subset across all experiments. We then repeat all experiments using initial weights pretrained to perform ImageNet classification. All tests are run five times with fixed hyperparameters to collect mean and standard deviation values. It is important to note that prior works do not report mean and standard deviation, likely resulting in inflated accuracy numbers. The reported results for AFFACT, for example, use the model which obtained the highest validation accuracy out of multiple runs.

For CelebA-C+A, we resize from the original $218 \times 178$ size to $274 \times 224$ to ensure the smallest dimension matches the $224 \times 224$ image size most commonly used for ImageNet. To augment images, we use flipping, cropping and rotation. Images are first resized by a random scale between 95% and 105%, then cropped back to $274 \times 224$. We then randomly rotate between $\pm 5$ degrees. We found that, while minor, the cropping and rotation transformations were useful for reducing overfitting. Finally, we flip the image horizontally with 50% probability. For CelebA-ITW, we zero-pad all images to be square then resize to $500 \times 500$ to ensure facial features remain visible even for images where the face is small. We then use the same augmentations adjusted to the larger image size. Because this increases the memory requirements of the network, we divide both the initial learning rate and batch size by 4.

To train our models, we primarily use the same parameters as the ResNet paper [10]: SGD with a batch size of 256, initial learning rate of 0.1, momentum of 0.9, weight decay of 0.0001, and a learning rate schedule in which the learning rate is multiplied by 0.1 when the validation loss plateaus. However, because we train for a fixed number of epochs, for most models we found that we obtained more consistent results by simply multiplying the learning rate by a factor of 0.9 every epoch. Exceptions include results without pretraining on our 10% downsampled versions of CelebA and on the full version of CelebA-ITW, for which we use the original plateau-based schedule. We also use a smaller multiplier of 0.8 for the pretrained model using all of CelebA-ITW. All models are trained on a single NVIDIA GTX 1080 Ti GPU. More detailed information about training hyperparameters is provided as supplemental material.

### B. Results

As shown in Table I, we are able to improve upon the CelebA-C+A results of MOON, CLMLE, and MCNN-AUX using ResNet-18 without any additional data, and, as shown in Table II, our CelebA-ITW results without additional data are within one standard deviation of all methods other than SA. Note that all methods which outperform our non-pretrained baselines use either auxiliary data or an additional model trained on a different dataset. AFFACT and AFFAIR use pre-trained face detectors, SSP+SSG and SA use semantic segmentation data, FAN uses semantic segmentation data

TABLE I

COMPARISON BETWEEN OUR BASELINE RESNET-18 NETWORKS AND STATE-OF-THE-ART METHODS ON THE CROPPED AND ALIGNED IMAGES (CELEBA-C+A). "10%" INDICATES THE NETWORK WAS TRAINED ON A SUBSET CONTAINING 10% OF THE TRAINING DATA.

| Method | Accuracy |
|---|---|
| MOON [21] | 90.94% |
| CLMLE [12] | 91.13% |
| MCNN-AUX [8] | 91.29% |
| AFFACT [6] | 91.67% |
| SSP + SSG [15] | 91.80% |
| FAN [9] | 91.81% |
| HFE [24] | 92.17% |
| PS-MCNN-LC [1] | 92.98 ± .25%[1] |
| ResNet-18 | 91.48 ± .06% |
| ResNet-18 (ImageNet pretrained) | 91.71 ± .01% |
| ResNet-18 (10%) | 90.32 ± .07% |
| ResNet-18 (10%, ImageNet pretrained) | 90.88 ± .02% |

TABLE II

COMPARISON BETWEEN OUR BASELINE RESNET-18 NETWORKS AND STATE-OF-THE-ART METHODS ON THE IN THE WILD IMAGES (CELEBA-ITW).

| Method | Accuracy |
|---|---|
| LNets+ANet [19] | 87% |
| Zhong et. al. [25] | 89.80% |
| AFFACT [6] | 91.45% |
| AFFAIR [18] | 91.45% |
| SA [16] | 91.47% |
| ResNet-18 | 91.36 ± .13% |
| ResNet-18 (ImageNet pretrained) | 91.81 ± .07% |
| ResNet-18 (10%) | 89.86 ± .13% |
| ResNet-18 (10%, ImageNet pretrained) | 90.43 ± .05% |

as well as ImageNet pretraining, and HFE and PS-MCNN use CelebA identity labels. Additionally, ResNet-18 has far fewer parameters and is much faster at inference time than the methods used in most other works. For example, SA uses an Inception-v3 backbone and AFFACT uses ResNet-50. Both networks have twice as many parameters as ResNet-18. Note that AFFACT reports higher accuracies when using 162 test-time augmentations or an ensemble of networks. For fairness of comparison we use their results using a single model and no test-time augmentations.

Notably, while AFFACT and AFFAIR use face detection or alignment transformations, we find that we are able to obtain high-quality results on CelebA-ITW without any alignment or face detection. With ImageNet pretraining, our results improve upon the nearest three methods, all of which are within 0.02% of each other, by 0.34%. We also improve upon our best CelebA-C+A results, despite most state-of-the-results using CelebA-C+A rather than CelebA-ITW. This is partially because the data was labeled using the original images, and some attributes are not visible in the aligned version. In particular, *wearing necklace* and *wearing necktie* are frequently cropped out of the aligned image. The full-size images may have also contributed to bias in the labeling which networks using aligned data cannot exploit.

[1][1] does not report the number of runs used to compute standard deviation.

Although the main advantage of CelebA is its large size, we are also able to obtain results comparable to state-of-the-art with just 10% of the training data available (a total of 16,277 training samples, rather than the 162,771 in the complete dataset). With ImageNet pretraining, our results for CelebA-C+A are competitive with MOON, which is used as the strongest baseline for accuracy comparison by several works [6][15][7]. All methods which improve upon our CelebA-ITW results, with or without ImageNet pretraining, use either a pretrained face detector or additional data.

## IV. IMPROVING EVALUATION

In this section we show that our ability to match or improve upon state-of-the-art using simple models is in part because currently used evaluation metrics are highly flawed. We provide suggestions for better evaluation and baselines and show that better metrics reveal labeling flaws which harm performance for many attributes.

### A. Better Metrics

Due to the imbalance present in CelebA, the accuracy of a model which always predicts the most common class based on the distribution of the training data is 79.91%, rather than 50% as it would be for a balanced dataset. For the least balanced attributes, such a model can obtain accuracy as high as 97.88%. To demonstrate why this is problematic for comparing different methods, we compare our baseline trained on 10% of the data with our baseline trained on the entire dataset. For the least balanced attributes, the network trained on 10% of the data only has a few hundred positive examples to learn from, so we expect these attributes to be where the difference between the two models is most apparent. However, when evaluating using accuracy, we observe the opposite: the most imbalanced attributes correspond to the smallest differences in accuracy. This is despite the fact that the network trained with less data clearly does worse on these attributes in terms of both precision and recall (combined using F1), as shown in Figure 2. Because there is little improvement than can be made over always predicting the majority class, performing well for highly

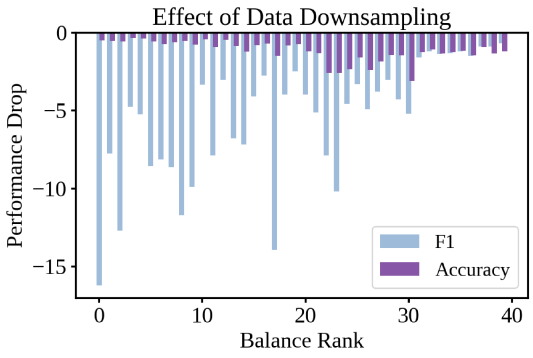


Fig. 2. Per-attribute accuracy and F1 drop incurred by training on a random selection of 10% of the training data. Balance rank orders attributes by their ratio between positive and negative samples, with rare attributes (e.g. *bald*) on the left and common attributes (e.g. *young*) on the right.

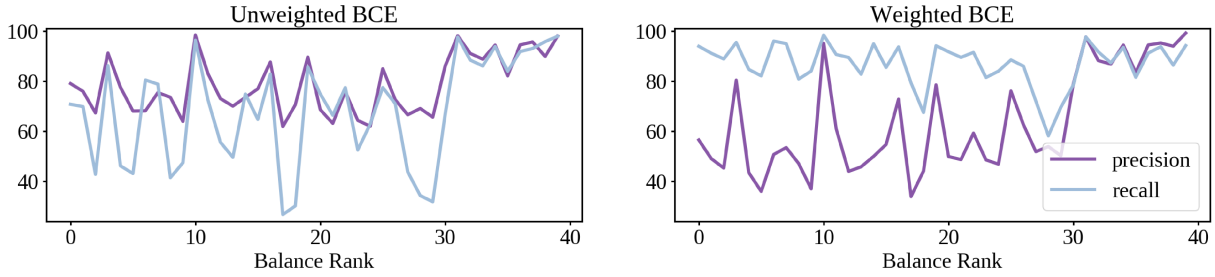


Fig. 3. Precision and recall for a ResNet-18 network optimized with BCE (left) and balance-weighted BCE (right). Balance rank orders attributes by their ratio between positive and negative samples, with rare attributes (e.g. *bald*) on the left and common attributes (e.g. *young*) on the right.

imbalanced attributes is not very important for achieving high average accuracy scores.

To address this issue, several previous works [13][4][15][9][12] have used balanced accuracy, which weighs true positive rate equally to true negative rate:

$$\frac{1}{2}\left(\frac{TP}{TP+FN}+\frac{TN}{TN+FP}\right) \tag{1}$$

Average precision has also been used [15]. However, we argue that both approaches are flawed. Average precision can be maximized at the expense of recall by only predicting 1 when highly confident, and balanced accuracy considers only true positive rate and true negative rate, which can be misleading for highly imbalanced data [3]. In the case of CelebA, balanced accuracy places very little weight on precision for highly imbalanced attributes. For the *bald* attribute, for example, a network can obtain a balanced accuracy of 95% with perfect recall but a precision of just 17.5% (2.12% TP, 10% FP, 87.78% TN, 0% FN). The relationship can be more clearly shown by rewriting balanced accuracy as the following:

$$\frac{1}{2}\left(1-\frac{FN}{N_p}\right)\left(1-\frac{FP}{N_n}\right), \tag{2}$$

where $N_p = TP+FN$ is the total number of positive samples and $N_n = TN+FP$ is the total number of negative samples. When $N_n >> N_p$, it is much more important to have few $FN$ than few $FP$, thus prioritizing recall. Similarly, when $N_p >> N_n$, it is much more important to have few $FP$ than few $FN$, thus prioritizing precision.

Because almost all CelebA attributes are predominately negative, we find in practice that optimizing models for balanced accuracy simply results in maximizing recall. To show this, we train our baseline ResNet-18 model using a loss function which is balanced by weighing each attribute loss by the ratio between negative and positive samples for that attribute. We find that our balanced accuracy on CelebA-C+A improves substantially with this weighted loss ($81.52 \pm .14$ to $87.84 \pm .11$), but this simply trades precision for recall, and our F1-score (the harmonic mean of precision and recall) remains unaffected ($71.99 \pm .23$ to $71.83 \pm .38$). The tradeoff between precision and recall is shown in Figure 3. Note that, particularly for the least balanced attributes, precision is substantially damaged to improve recall.

To further show how balanced accuracy can be problematic, we replace each attribute weight $w$ for our BCE loss with $w^{1.5}$, thus improving recall for attributes which are mostly negative and improving precision for attributes which are mostly positive. With ImageNet pretraining, we obtain a balanced accuracy of $88.43 \pm .05\%$ – just 0.35% below the state-of-the-art result obtained by CLMLE using DeepID2 features pretrained to perform verification on CelebA – while our average precision drops from $78.75 \pm .13\%$ to $58.62 \pm .10\%$ and our accuracy drops from $91.72 \pm .01\%$ to $86.10 \pm .12\%$.

While using a combination of balanced accuracy, accuracy, and average precision overcomes their collective issues, this can lead to practical difficulties in comparing models. For example, [12] compares their model to [15] using both balanced accuracy and accuracy. However, Kalayeh et. al. obtained their results using two separate models, one optimized for accuracy and the other optimized for balanced accuracy, thus limiting the usefulness of the comparison.

In light of these results, we argue that all metrics used by prior work – accuracy, balanced accuracy, and average precision – are insufficient for measuring attribute prediction performance, particularly for imbalanced attributes. We instead suggest F1-score, which is commonly used for other problems exhibiting class imbalance [14] and avoids the problems described above by completely ignoring the number of true negatives. We provide results for all our models using accuracy, balanced accuracy, and F1 in Table III. Following previous work, our balanced accuracy metrics are computed separately for each attribute then averaged to enforce that the model should do well for all attributes. Similarly, F1 results are the average of per-attribute F1 scores because using cumulative TP/FP/FN counts across all attributes result in prioritizing attributes which have more total positives. Note that our F1 results are therefore not comparable to those provided by [9], as they do not compute F1 separately for each attribute but instead use cumulative counts. To our knowledge, no other work provides F1 results for CelebA.

### B. Better Baselines

Given the small scale of accuracy differences between state-of-the-art approaches, it is worth considering how large of an effect hyperparameters and network backbone selection

TABLE III

ACCURACY, BALANCED ACCURACY, AND F1 SCORES AVERAGED OVER ALL ATTRIBUTES.

| | Without Pretraining | | | With ImageNet Pretraining | | |
|---|---|---|---|---|---|---|
| Data | Acc. | Bal. Acc. | F1 | Acc. | Bal. Acc. | F1 |
| CelebA-C+A, 100% | 91.48 ± .06% | 81.49 ± .18% | 71.98 ± .21% | 91.71 ± .01% | 82.35 ± .11% | 73.19 ± .10% |
| CelebA-C+A, 10% | 90.32 ± .07% | 78.20 ± .43% | 66.58 ± .76% | 90.88 ± .02% | 79.62 ± .12% | 69.05 ± .11% |
| CelebA-ITW, 100% | 91.36 ± .13% | 82.80 ± .11% | 73.13 ± .21% | 91.81 ± .07% | 82.39 ± .25% | 73.36 ± .31% |
| CelebA-ITW, 10% | 89.86 ± .13% | 76.34 ± .72% | 63.72 ± 1.58% | 90.43 ± .05% | 76.92 ± .19% | 64.57 ± .28% |

can make. We find that several seemingly minor changes in training hyperparamters can result in substantial differences in validation accuracy. For example, we found that resizing images from $218 \times 178$ to $274 \times 224$ resulted in an average validation accuracy improvement of 0.41%. Additionally, the final model after our fixed number of epochs does not always achieve the highest validation accuracy, and stopping training early can result in similar gains. In particular, the learning rate reduction on plateau schedule varies a large amount from run to run, and can result in standard deviations as high as 0.13 as shown in Table III. These fluctuations highlight the importance of having a strong, directly comparable baseline to show that reported improvements are actually a result of the proposed method.

This can be further seen from works which provide such a comparison. [16] use Inception-v3 as the backbone of their proposed Symbiotic Augmentation (SA), and as such provide comparisons to an Inception-v3 baseline trained without SA. Their method only improves upon this baseline by 0.15%. Another method which uses segmentation masks, [9], achieves accuracy results for CelebA-C+A which are within 0.01% of SSP+SSG (the precursor to SA), but improve upon their ResNet-50 baseline by a much larger 0.31%. While the lack of mean and standard deviation numbers for these results makes it difficult to determine how significant the improvements over these baselines are, it is clear from their small scale that a large portion of the difference between methods comes from backbone networks and hyperparameter selection.

### C. Labeling Inconsistencies

The similarity in accuracy between a classifier trained using all the data and one trained using just 10% of the data raises the question of why state-of-the-art classifiers struggle to obtain accuracy greater than 92%. Measuring results in terms of F1 demonstrates some of the major problems present in the dataset. As explored by previous work, several labels such as *oval face*, *attractive*, *high cheekbones*, and *arched eyebrows* are subjective and inconsistently labeled, while other labels, such as *lipstick*, are frequently mislabeled [7]. While some methods have been able to obtain good results on these attributes in terms of accuracy or balanced accuracy by exploiting the balance of the dataset, when measured in terms of F1 these issues become far more clear. For certain highly subjective attributes such as *narrow eyes*, *oval face*, and *big lips*, our baseline model pretrained using ImageNet is unable to obtain an F1-score above 50. For some of these attributes, the labeling issues can be seen

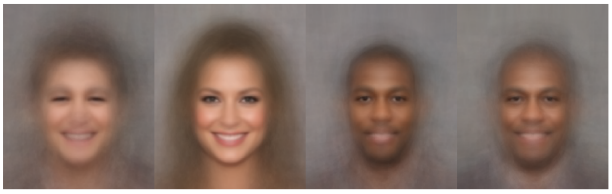

Fig. 4. Average of the 200 validation images which achieve the highest activations for *narrow eyes*, *high cheekbones*, *big lips*, and *big nose* (left to right).

by averaging the 200 validation images which result in the highest activations for each attribute. For example, as shown in Figure 4, *narrow eyes* frequently applies to partially closed eyes due to laughing and *high cheekbones* seems to just recognize smiling (we find that 85.6% of images labeled with *high cheekbones* are also labeled with *smiling*). For some subjective attributes such as *big nose* and *big lips*, our baseline seems to rely heavily on racial or gender bias. Of the 200 images with the highest activations for *big lips*, 99% of the people are black. Of the images with the lowest activations, 0% are black. For the top activations for *big nose*, 78% of the people are black and 3% are female. For the lowest activations, 0% are black and 100% are female. Almost half (46.5%) of the images that achieve the highest 200 activations for *big lips* are also in the highest 200 activations for big nose, suggesting that the two features learn similar biases.

Due to the subjectivity of these attributes, even when optimizing for accuracy rather than a balanced metric, the most balanced attributes aren't necessarily the ones the network performs best for. For example, as shown in Table IV, *big lips*, *oval face*, and *pointy nose* are among the most balanced attributes in the training set with a positive/negative ratio near 30, but we are unable to obtain an F1 much better than 50 for any. Additionally, our ability to obtain better performance on CelebA-ITW than any model not using identity labels on either CelebA-ITW or CelebA-C+A suggests that the labels are affected by factors outside of facial features. As previously mentioned, the *wearing necklace* attribute is frequently not visible in aligned images, allowing our ImageNet-pretrained network trained on CelebA-ITW to obtain an average F1 improvement of 15.28 over an identical network trained on CelebA-C+A. While the network trained using CelebA-C+A performs better on most attributes, there are 13 other attributes for which the network using unaligned data performs better, including *oval face* (+9.95), *mustache* (+5.46), *big lips* (+1.96), and *wearing necktie* (+1.92). While *wearing necktie* is more visible in the uncropped data,

TABLE IV

AVERAGE ACCURACY, BALANCED ACCURACY, PRECISION, RECALL, AND F1 RESULTS ON CELEBA-C+A USING OUR BASELINE RESNET MODEL WITH IMAGENET PRETRAINING. "%POS" IS THE PERCENTAGE OF SAMPLES WHICH ARE POSITIVE. WE BOLD ATTRIBUTES WITH AN F1 BELOW 60.

| Attribute | %pos | Acc | BA | Prc | Rcl | F1 |
|---|---|---|---|---|---|---|
| No Beard | 85.4 | 96.5 | 93.4 | 98.1 | 97.8 | 98.0 |
| Young | 75.7 | 89.0 | 82.6 | 90.9 | 95.1 | 92.9 |
| Wearing Lipstick | 52.1 | 94.2 | 94.2 | 95.9 | 92.8 | 94.3 |
| Smiling | 50.0 | 93.4 | 93.4 | 94.7 | 91.9 | 93.3 |
| Attractive | 59.6 | 83.2 | 83.2 | 83.8 | 82.0 | 82.8 |
| Mouth Slightly Open | 49.5 | 94.3 | 94.3 | 94.9 | 93.6 | 94.2 |
| High Cheekbones | 48.2 | 88.1 | 88.0 | 89.3 | 85.5 | 87.4 |
| Heavy Makeup | 40.5 | 92.0 | 91.5 | 91.5 | 88.5 | 90.0 |
| Male | 38.7 | 98.4 | 98.2 | 98.3 | 97.5 | 97.9 |
| Wavy Hair | 36.4 | 85.3 | 82.0 | 87.2 | 69.9 | 77.6 |
| **Big Lips** | 32.7 | 72.8 | 62.7 | 67.1 | 33.3 | 44.5 |
| **Oval Face** | 29.6 | 76.1 | 63.0 | 72.3 | 31.1 | 43.5 |
| **Pointy Nose** | 28.6 | 77.8 | 68.2 | 66.1 | 45.9 | 54.2 |
| Arched Eyebrows | 28.4 | 84.4 | 80.8 | 72.4 | 72.7 | 72.5 |
| Black Hair | 27.2 | 90.5 | 86.5 | 85.9 | 77.7 | 81.6 |
| Big Nose | 21.2 | 84.3 | 76.8 | 62.8 | 63.8 | 63.3 |
| Straight Hair | 21.0 | 85.0 | 74.7 | 66.7 | 56.9 | 61.4 |
| Wearing Earrings | 20.7 | 90.7 | 86.2 | 76.8 | 78.6 | 77.7 |
| Bags Under Eyes | 20.3 | 85.5 | 78.8 | 63.4 | 67.5 | 65.3 |
| Brown Hair | 18.0 | 89.5 | 83.7 | 69.3 | 74.6 | 71.9 |
| Bangs | 15.6 | 96.2 | 91.9 | 89.4 | 85.8 | 87.6 |
| **Narrow Eyes** | 14.9 | 87.7 | 64.5 | 69.4 | 31.5 | 43.3 |
| **Wearing Necklace** | 13.8 | 88.1 | 65.2 | 63.3 | 33.5 | 43.8 |
| Blond Hair | 13.3 | 96.2 | 91.4 | 86.3 | 84.8 | 85.5 |
| Bushy Eyebrows | 13.0 | 93.0 | 80.3 | 78.9 | 63.2 | 70.2 |
| 5 o Clock Shadow | 10.0 | 94.8 | 87.0 | 72.7 | 77.1 | 74.8 |
| **Receding Hairline** | 8.5 | 94.0 | 74.9 | 69.9 | 51.9 | 59.6 |
| Rosy Cheeks | 7.2 | 95.4 | 77.4 | 73.5 | 56.3 | 63.8 |
| Wearing Necktie | 7.0 | 97.1 | 87.1 | 81.7 | 75.4 | 78.4 |
| Eyeglasses | 6.5 | 99.7 | 98.4 | 98.2 | 96.9 | 97.5 |
| **Chubby** | 5.3 | 95.9 | 75.7 | 64.2 | 53.1 | 58.1 |
| **Blurry** | 5.1 | 96.4 | 73.4 | 71.2 | 47.9 | 57.2 |
| Sideburns | 4.6 | 98.0 | 89.6 | 76.6 | 80.4 | 78.5 |
| Goatee | 4.6 | 97.6 | 88.1 | 72.0 | 77.7 | 74.7 |
| **Double Chin** | 4.6 | 96.5 | 72.7 | 66.2 | 46.6 | 54.7 |
| Pale Skin | 4.2 | 97.2 | 75.3 | 74.2 | 51.4 | 60.7 |
| Wearing Hat | 4.2 | 99.2 | 94.3 | 91.2 | 89.0 | 90.1 |
| **Mustache** | 3.9 | 97.1 | 72.7 | 68.8 | 46.3 | 55.4 |
| Gray Hair | 3.2 | 98.3 | 84.9 | 74.6 | 70.7 | 72.5 |
| Bald | 2.1 | 99.1 | 86.8 | 80.4 | 74.1 | 77.1 |

Fig. 5. Top row: Examples of validation images labeled as *bald*. Some images are clearly not bald (leftmost example) or clearly bald (rightmost example), but there is some ambiguity in between. Bottom row: examples of validation images labeled as *receding hairline* seemingly due to close-cropped or tied back hair.

TABLE V

AVERAGE FLEISS $\kappa$ AGREEMENT FOR THE 40 ATTRIBUTES IN CELEBA. $\kappa < 0$ INDICATES AGREEMENT IS WORSE THAN WOULD BE EXPECTED BY RANDOM CHANCE, $\kappa = 1$ INDICATES PERFECT AGREEMENT.

| Attribute | $\kappa$ |
|---|---|
| Blurry | −0.0181 |
| Pale Skin | 0.1562 |
| Mouth Slightly Open | 0.2141 |
| Narrow Eyes | 0.2378 |
| Wearing Hat | 0.2489 |
| Smiling | 0.2551 |
| Wearing Necktie | 0.2712 |
| Double Chin | 0.2910 |
| Wearing Necklace | 0.3113 |
| High Cheekbones | 0.3178 |
| Rosy Cheeks | 0.3282 |
| Bags Under Eyes | 0.3388 |
| Receding Hairline | 0.3405 |
| Straight Hair | 0.3441 |
| Brown Hair | 0.3719 |
| Oval Face | 0.3881 |
| Wearing Earrings | 0.3893 |
| Chubby | 0.3924 |
| Eyeglasses | 0.4150 |
| Gray Hair | 0.4233 |

| Attribute | $\kappa$ |
|---|---|
| Wavy Hair | 0.4272 |
| Bangs | 0.4313 |
| Pointy Nose | 0.4546 |
| Black Hair | 0.4717 |
| Sideburns | 0.4759 |
| Bushy Eyebrows | 0.4893 |
| Mustache | 0.4945 |
| Bald | 0.4981 |
| Goatee | 0.5062 |
| 5 o Clock Shadow | 0.5131 |
| Arched Eyebrows | 0.5131 |
| Attractive | 0.5140 |
| Big Nose | 0.5585 |
| Blond Hair | 0.5727 |
| Heavy Makeup | 0.6302 |
| No Beard | 0.6450 |
| Big Lips | 0.7279 |
| Wearing Lipstick | 0.7322 |
| Young | 0.8360 |
| Male | 0.9789 |

the other features should be entirely visible in the aligned images and may therefore be biased by factors cropped out during alignment.

We also find that there are many attributes other than those described in [7] which are seemingly non-subjective but lack clear definitions and are inconsistently labeled. For example, we found it highly unclear what differentiates *bald* from *receding hairline*. Although the two classes should seemingly be disjoint, 33.1% of images labeled with *bald* are also labeled with *receding hairline*. Though detailed analysis of labeling issues is left for future work, we found that as many as 50% of images labeled as bald have some amount of hair on the scalp. *Receding hairline* is even more inconsistently labeled, with labelers frequently seeming to use it to describe hair which is close-cropped or tied back. Examples of both are shown in Figure 5. More samples of these attributes are provided as supplemental material.

For a more complete evaluation of consistency, we use the Fleiss' $\kappa$ measure commonly used for evaluating inter-rater agreement [5]. Fleiss' $\kappa$ is defined as $\frac{\bar{P}-P_e}{1-P_e}$, where $\bar{P}$ is the probability that two randomly selected reviewers agree on a specific rating for a subject and $P_e$ is the probability that this would occur by chance (if an attribute is very rare, for example, agreement would be high even if the reviewers are using entirely different criteria simply because most ratings would be 0). Attributes such as *oval face* and *pointy nose* should be consistent for different images of the same person, so labels for different images of the same person can be considered different ratings of the same subject for the purpose of computing $\kappa$. As shown in Table V, we find that agreement is poor for most attributes. In fact, many of the highest agreement scores are for attributes which we expect to be *least* consistent across different images, such as *wearing lipstick*, *heavy makeup*, and *5 o' clock shadow*. This lack of consistency may explain why using CelebA attributes

have not shown significant value for face verification. For example, [21] found that using their CelebA attribute classifier for face verification only moderately improved over a similar results from Kumar et. al. in 2009, which used more attributes but far less training data. Both methods lag far behind state-of-the-art approaches. It is possible that facial attributes may be more useful for verification than these results indicate, but only if better care is taken during the labeling process to ensure attributes are labeled consistently.

## V. CONCLUSION

Although CelebA is the largest-scale facial attribute dataset available, it is difficult to directly compare methods trained on this data. The two metrics primarily used to compare performance, accuracy and balanced accuracy, can be optimized for imbalanced attributes without producing a classifier which is actually useful for predicting those attributes. We demonstrate that simple baseline models are able to obtain results very close to highly specialized methods. To our knowledge, no method is able to improve upon a non-pretrained ResNet-18 model without requiring additional data or an additional pretrained model, and improvements over a ResNet-18 model pretrained on ImageNet are small (or, in the case of the uncropped data, nonexistent). Additionally, many attributes have highly inconsistent or inaccurate labels, making it difficult for any model to achieve reasonable results. Note that LFWA, a popular alternative to CelebA, was labeled by the same group in the same manner as CelebA and as such suffers from similar issues.

To improve evaluation of facial attribute prediction models, we suggest using metrics which are invariant to true negative count such as F1, computed as the average of per-attribute scores to ensure that all attributes are weighed evenly. Per-attribute results showing which attributes the model performs best on are also important both to show how performance is impacted by balance and to demonstrate which attributes cannot be reliably predicted. Improved performance on certain poorly-labeled attributes may not be meaningful. Additionally, due to the relatively small differences between most methods and the varying use of additional data, we emphasize the importance of comparing to strong baselines and providing mean and standard deviation numbers to ensure reported improvements come from the proposed method rather than hyperparameter and backbone selection. We hope these suggestions will improve the community's ability to evaluate new methods for facial attribute prediction.

## REFERENCES


[1] J. Cao, Y. Li, and Z. Zhang. Partially shared multi-task convolutional neural network with local constraint for face attribute learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4290–4299, 2018.

[2] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[3] J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

[4] Q. Dong, S. Gong, and X. Zhu. Class rectification hard mining for imbalanced deep learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1851–1860, 2017.

[5] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[6] M. Günther, A. Rozsa, and T. E. Boult. Affact: Alignment-free facial attribute classification technique. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 90–99. IEEE, 2017.

[7] E. Hand, C. Castillo, and R. Chellappa. Doing the best we can with what we have: Multi-label balancing with selective learning for attribute prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[8] E. Hand and R. Chellappa. Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[9] K. He, Y. Fu, W. Zhang, C. Wang, Y.-G. Jiang, F. Huang, and X. Xue. Harnessing synthesized abstraction images to improve facial attribute recognition. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 733–740. International Joint Conferences on Artificial Intelligence Organization, 7 2018.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 28(11):5464–5478, 2019.

[12] C. Huang, Y. Li, C. L. Chen, and X. Tang. Deep imbalanced learning for face recognition and attribute prediction. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[13] C. Huang, Y. Li, C. C. Loy, and X. Tang. Learning deep representation for imbalanced classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[14] J. M. Johnson and T. M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):27, 2019.

[15] M. M. Kalayeh, B. Gong, and M. Shah. Improving facial attribute prediction using semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6942–6950, 2017.

[16] M. M. Kalayeh and M. Shah. On symbiosis of attribute prediction and semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[17] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *2009 IEEE 12th international conference on computer vision*, pages 365–372. IEEE, 2009.

[18] J. Li, F. Zhao, J. Feng, S. Roy, S. Yan, and T. Sim. Landmark free face attribute prediction. *IEEE Transactions on Image Processing*, 27(9):4651–4662, 2018.

[19] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[20] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa. An all-in-one convolutional neural network for face analysis. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 17–24. IEEE, 2017.

[21] E. M. Rudd, M. Günther, and T. E. Boult. Moon: A mixed objective optimization network for the recognition of facial attributes. In *European Conference on Computer Vision*, pages 19–35. Springer, 2016.

[22] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. *Advances in neural information processing systems*, 27:1988–1996, 2014.

[23] Z. Wang, K. He, Y. Fu, R. Feng, Y.-G. Jiang, and X. Xue. Multi-task deep neural network for joint face recognition and facial attribute prediction. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 365–374, 2017.

[24] J. Yang, J. Fan, Y. Wang, Y. Wang, W. Gan, L. Liu, and W. Wu. Hierarchical feature embedding for attribute recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13055–13064, 2020.

[25] Y. Zhong, J. Sullivan, and H. Li. Leveraging mid-level deep representations for predicting face attributes in the wild. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3239–3243. IEEE, 2016.