University of Nevada, Reno

Face Captioning Using Prominent Feature Recognition

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science in
Computer Science and Engineering

by

Bryson Lingenfelter

Dr. Emily M. Hand - Thesis Advisor
May, 2021

THE GRADUATE SCHOOL

We recommend that the thesis
prepared under our supervision by

Bryson Lingenfelter

entitled

Face Captioning Using Prominent Feature Recognition

be accepted in partial fulfillment of the
requirements for the degree of

Master of Science

Emily M. Hand
Advisor

George Bebis
Committee Member

Floris van Breugel
Graduate School Representative

David W. Zeh, Ph.D., Dean
Graduate School

May, 2021

Abstract

Humans rely on prominent feature recognition to correctly identify and describe previously seen faces. Despite this fact, there is little existing work investigating how prominent facial features can be automatically recognized and used to create natural language face descriptions. Facial attribute prediction, a more commonly studied problem in computer vision, has previously been used for this task. However, the evaluation metrics and baseline models currently used to compare different attribute prediction methods are insufficient for determining which approaches are best at classifying highly imbalanced attributes. We also show that CelebA, the largest and most widely used facial attribute dataset, is too poorly labeled to be suitable for prominent feature recognition. To deal with these issues, we propose a method for generating weak prominent feature labels using semantic segmentation and show that we can use these labels to improve attribute-based face description.

Acknowledgments

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

Table of Contents

1 Introduction **1**

2 Background **4**

2.1 Prominent Features 4

2.2 Attribute Recognition 6

2.2.1 Attribute Recognition Networks 6

2.2.2 Labeling Issues 7

2.3 Captioning 8

3 Improving Attribute Recognition **10**

3.1 Improving Evaluation 10

3.1.1 Baseline Experiments 12

3.1.2 Improving Evaluation 16

3.1.3 Summary 25

3.2 Data Issues 26

3.2.1 Incorrect Labels 26

3.2.2 Inconsistent Labels 30

3.2.3 Summary 35

4 Utilizing Attributes for Face Captioning **37**

4.1 Weakly Labeled Prominent Features 37

4.1.1 Data 38

4.1.2 Weak Labeling 39

4.1.3 Evaluation 42

4.1.4 Attribute Prediction 44

4.2 Captioning 45

4.2.1 Implementation 46

4.2.2 Evaluation 48

4.2.3 Summary 52

5 Conclusion and Future Work 54

List of Tables

3.1 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. 15

3.2 Comparison between our baseline ResNet-18 networks and state-of-the-art methods on the in the wild images (CelebA-ITW). 15

3.3 Accuracy, balanced accuracy, and F1 scores averaged over all attributes. 19

3.4 Average accuracy, balanced accuracy, precision, recall, and F1 results on CelebA-C+A using our baseline ResNet model with ImageNet pre-training. "%pos" is the percentage of samples which are positive. We bold attributes with an F1 below 60. 24

3.5 Contradicting attribute labels. % is the percentage of images in the full dataset with the label in the left column and a contradicting label from the middle column. 27

3.6 Number of false positives (FP) and false negatives (FN) for non-subjective attributes out of a sample of 400. Gender labels were verified using identity labels. 29

3.7 F1 scores for linear classifiers using estimated quantitative metrics to predict subjective labels compared with a ResNet-18 classifier trained on CelebA. All attributes other than *young* are evaluated using only frontal images. 32

3.8 Average Fleiss κ agreement for the 40 attributes in CelebA 34

4.1 Shape measurement computation details. All metrics are normalized by distance between the eyes unless otherwise specified 41

4.2 Entropy, information gain, and the ratio between the two for each feature/label combination in the caricature dataset. 43

4.3 Comparison between randomly-generated captions in Multi-Modal CelebA (left) and captions in our datast (right) for the first five images in CelebA-HQ. 47

4.4 Verification accuracy for CelebA and shape label-based captions using consistent and inconsistent labels. 49

4.5 Validation accuracy for CelebA and shape label-based captions using consistent and inconsistent labels. 51

List of Figures

1.1 Examples of CelebA-ITW images (top) and their CelebA-C+A versions (bottom). 2

3.1 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. 17

3.2 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. 18

3.3 Average of the 200 validation images which achieve the highest activations for *narrow eyes*, *high cheekbones*, *big lips*, and *big nose* (left to right). 22

3.4 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. 23

3.5 The first four images in CelebA labeled as *double chin* but not *chubby*. None of these images contain a double chin. 27

3.6 Examples of false positives (left to right): *eyeglasses*, *wearing earrings*, *wearing hat*, and *wearing necktie*. 29

3.7 Left: Histogram of combined lip segment size for images labeled with *big lips* and images not labeled with *big lips*. Right: Histogram of estimated ages for images labeled with *young* and images not labeled with *young*. 31

4.1 Examples of two subjects in the caricature set and their associated ground-truth labels. 38

4.2 Visualization of a segmentation mask output by the DeepLab v3+ network. 40

4.3 Plots of an $ax+b$ function fit eye contours (left) and an $alog(x)+bx+c$ function fit to eyebrow contours (right). 40

4.4 Examples of images at different orientations rotated using Rotate-and-Render 45

4.5 Application for searching for images using text descriptions. The selected images are pulled from a pool of 2,000 validation images. . . . 52

Chapter 1

Introduction

Research in human vision indicates that humans recognize faces by identification of prominent features [1], and in some cases can even recognize caricatures (exaggerated drawings of faces) more quickly than unaltered photos of faces [2]. These features are important for generating useful descriptions of faces. However, there is little existing work in computer vision investigating how best to recognize prominent features and even less work in natural language processing investigating how to use these features to describe faces with words. This thesis aims to address both issues.

Prior work in prominent facial feature recognition exists in the form of facial attribute prediction. 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 [3–5] semantic segmentation [6], and other face parsing tasks such as detection and landmarking [7]. Facial attributes have also recently become popular for GAN-based face editing [8, 9]. The only large-scale facial attribute dataset that includes prominent feature descriptions is CelebA [10], which contains $202,599$ images of $10,177$ people labeled with 40 binary attributes.

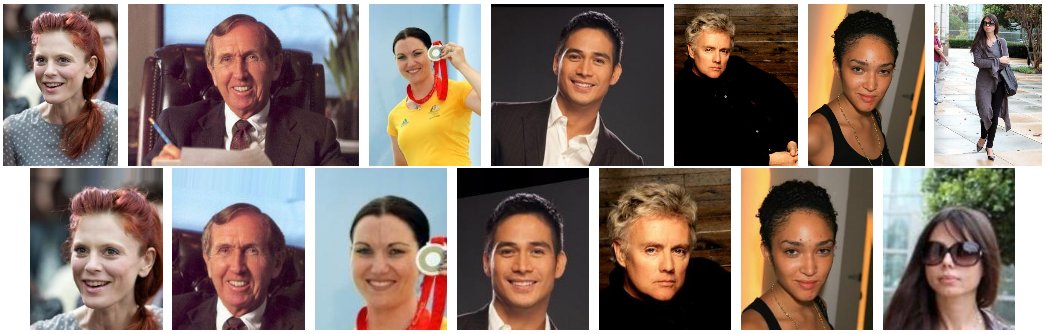


Figure 1.1: Examples of CelebA-ITW images (top) and their CelebA-C+A versions (bottom).

The images are provided in both the original, uncropped format and as 218 × 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 1.1.

In Chapter 3, we investigate the applicability of existing attribute prediction methods to prominent feature recognition. We find that the CelebA dataset is largely unsuitable for prominent feature recognition and that there exist a large number of flaws with existing methods for model evaluation. We are able to train simple ResNet-18 models to achieve state-of-the-art or near state-of-the-art attribute prediction performance without taking advantage of any assumptions specific to the domain. We show that this is because existing works fail to effectively evaluate how well their models perform for imbalanced attributes, which are essential for prominent feature recognition because prominent features are necessarily uncommon. We then argue that these issues with evaluation are partly caused by larger issues with the dataset itself. We show that many attributes in CelebA are frequently labeled incorrectly, contradict with other labels, or are highly inconsistent with respect to quantitative measurements. In light of these results, we argue that CelebA and other existing attribute datasets cannot directly be used for prominent feature recognition. To improve these issues, we suggest better evaluation metrics and awareness of data issues to facilitate more useful methods for attribute prediction.

While there are many issues with CelebA, it remains the primary dataset used for facial attribute prediction. There has recently been increasing interest in using text data to manipulate face images, and as such some works have used CelebA attributes to generate captions for GAN training [11,12]. Captions are also useful for improving accessibility and for image lookup, as existing image description methods do not describe faces in a way that identifies what differentiates the person from other individuals. Despite these applications, there has been little work focusing on generating useful face captions.

In Chapter 4, we suggest alternative methods for prominent feature recognition and use prominent features to generate natural language descriptions of face images. Because there does not exist prior work for evaluating face captions, we also propose a framework for evaluating how well a caption describes what is unique about a face. We use this to determine the quality of captions generated using facial attribute labels. To improve upon these labels, we also propose a novel method for weakly-labeled prominent feature detection. We collect a small dataset of unrestricted prominent feature descriptions and use feature sizes and shapes estimated using semantic segmentation to predict many of these features. We then train a classifier to predict these weak labels and show that we can use the predicted features to improve the quality of our generated captions. We also show how face captions can be used in practice for image lookup.

By examining issues in existing face attribute data and proposing empirical methods for evaluating description quality, we hope to facilitate future research in natural language face description. In Chapter 5 we provide discussion of the challenges present in face captioning and describe several potential avenues for future work in the area.

Chapter 2

Background

2.1 Prominent Features

Research in human vision has found that prominent facial features are essential for face identification, and that humans may have evolved to have more variable facial morphology to assist quick recognition [1]. Humans can in some cases even recognize caricatures, drawings which exaggerate the most prominent features of a face, more quickly than actual face photos [2]. Existing work has also found that humans can more easily recognize a sketch of a face when the prominent features are exaggerated, but have more difficulty with recognizing sketches when prominent features are made less distinct [13]. Humans have more difficulty recognizing face photos which have been distorted to be more similar to an average face, but may even be able to recognize photos distorted to exaggerate distinctions from an average face more easily than a non-distorted image [14].

Because prominent features are important for human vision, some existing works have used visual features for computer vision identification and verification models.

Facial attributes were originally proposed as a method for using describable features to improve face verification [3] and allow for text-based face image lookup [15]. This method used labels collected through Amazon Mechanical Turk, with binary labels for each feature. Features were selected arbitrarily, and are not useful for describing all people (e.g., race labels are included, but the only options are "white," "black," and "Asian"). A much larger dataset, CelebA [10], was constructed using a subset of these attributes. We discuss this data and models for predicting its labels in Section 2.2.

While there exist other facial attribute datasets, we focus on CelebA because it is the largest and most widely used. PubFig [3] contains 65 attributes, but only 1,000 samples per attribute were collected and the dataset only contains 200 unique identities. LFWA [10] contains 40 attributes for 13,233 images, but was labeled by the same group and in the same manner as CelebA, which contains 202,599 images. Other attribute datasets such as Adience [16] and IMBD-Wiki [17] typically only contain a small number of attributes which cannot be used for prominent features such as race, gender, or age.

Prominent features can also be detected by modeling the shape of the face and directly computing what features are prominent rather than predicting binary attributes. This can be done with either semantic segmentation, which annotates each pixel with a label indicating which features it is part of (e.g., face, nose, lips, eyes, etc.) [18, 19] or 3D modeling, which builds a complete predicted 3D face mesh for a given image [20,21]. To our knowledge no prior work has investigated the applicability of these methods to prominent feature recognition.

2.2 Attribute Recognition

2.2.1 Attribute Recognition Networks

The CelebA dataset contains the following 40 attributes: 5 o'clock shadow, arched eyebrows, attractive, bags under eyes, bald, bangs, big lips, big nose, black hair, blond hair, blurry, brown hair, bushy eyebrows, chubby, double chin, eyeglasses, goatee, gray hair, heavy makeup, high cheekbones, male, mouth slightly open, mustache, narrow eyes, no beard, oval face, pale skin, pointy nose, receding hairline, rosy cheeks, sideburns, smiling, straight hair, wavy hair, wearing earrings, wearing hat, wearing lipstick, wearing necklace, wearing necktie, and young. Each image in the dataset contains a binary label for each attribute indicating whether the attribute is present or not. Models trained on the dataset therefore take an image as input an predict 40 binary labels as output.

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 CNNs, LNet_0, 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 [10]. This was the first proposed method designed specifically for CelebA attribute prediction and obtained significantly better results than the previous attribute recognition methods PANDA [22] and FaceTracer [23].

Later works rely on more typical end-to-end CNN models. MOON [24], 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 [25], 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 [26] uses a shallower CNN with different branches for different attribute groupings to take advantage of relationships between attributes.

Other works use additional data to improve performance. SSP+SSG [27] 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 [6]. Segmentation data has also been used by [28], who use a 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 [29] 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 [30], which was used to create CelebA, effectively meaning that CelebA identity labels are auxiliary data. HFE [31] 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 and obtains, to our knowledge, the best reported results on aligned CelebA.

2.2.2 Labeling Issues

While many previous works have used the CelebA dataset to evaluate attribute prediction and imbalanced classification methods, few have provided analysis of labeling issues. Hand et. al. [32] argue that the poor performance of state-of-the-art classifiers on many attributes is caused by ambiguous labeling, and provide examples of poor labels for the attributes *oval face*, *attractive*, *high cheekbones*, and *arched eyebrows*. They also show that many images labeled with *lipstick* are incorrectly labeled. There has also been some work discussing the bias caused by subjective labeling and dataset imbalance. Prabhu et. al. [33] show that increasing the contribution of labels such as *attractive* and *wearing lipstick* to a generative model causes images to look like blond, white women. Wang et. al. [34] show that the imbalance present in CelebA results in classifiers amplifying bias. Other works have shown that bias amplification is an issue in large scale datasets exhibiting imbalance [35]. However, to our knowledge no other work has performed quantitative analysis of labeling issues in CelebA. In section 3.2 we show that many CelebA labels, in addition to being subjective and imbalanced as shown in prior work, are frequently inconsistent or even completely incorrect.

2.3 Captioning

Existing work in image captioning typically uses encoder-decoder models with large datasets of image-caption pairs and recurrent or transformer-based decoders for generating text [36, 37]. While there are large existing caption datasets for birds [38], common objects [39] and surveillance imagery [40], to our knowledge the only face caption dataset with human annotation is Face2Text [41], which only provides captions for 400 face images with very little direction for how labelers should describe faces. Although the dataset collectors filtered the captions for “hate speech,” this only resulted in one description being discarded and descriptions which contain “use of ethnic or other characteristics when these are not used in an offensive manner” were left in the dataset. Most other work dealing with face captions focuses on extracting names to collect additional data for face verification [42, 43] rather than collecting descriptions of face features.

As a result of these data issues, it is difficult to apply supervised learning to face captioning. While there is some existing work in unsupervised captioning [44], this relies on a large corpus of unpaired caption texts. Unfortunately, captions for face images are very different from captions for common objects due to the smaller amount of variation, so it is far more difficult to construct an adequate dataset. As a result, existing works using face caption have relied upon captions generated directly from attribute labels using a set of pre-defined rules [12,45]. Prior works have used a probabilistic context free grammar (PCFG) to generate captions using CelebA attribute labels [11, 12]. However, to our knowledge no prior work has investigated how best to generate and evaluate face captions or use feature descriptions other than those provided with CelebA.

Chapter 3

Improving Attribute Recognition

In this chapter we investigate issues with attribute recognition and the CelebA attribute dataset. Section 3.1 deals with evaluation issues present in existing attribute recognition techniques, and section 3.2 deals with issues with the dataset itself.

3.1 Improving Evaluation

In this section we show that near-state-of-the-art accuracy can be obtained on both versions of the CelebA dataset using a ResNet-18 model [46] 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 pretraining 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.

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. We provide our implementation and detailed per-attribute results (to be made publicly available following publication) as a simple but strong baseline for future work to compare to.

3.1.1 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 Chapter 2 following this approach, even when using far less data.

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 218×178 size to 274×224 to ensure the smallest dimension matches the 224×224 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 274×224. We then randomly rotate between ± 5 and 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 500×500 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 [46]: 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. Full configurations for each experiment are described below.

- 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

Results

As shown in Table 3.1, we are able to improve upon the CelebA-C+A results of MOON and CLMLE using ResNet-18 without any additional data, and, as shown in Table 3.2, 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 as well as ImageNet pretraining, and HFE uses 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, which contains twice as many parameters as ResNet-18. AFFACT uses ResNet-50, which similarly has 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 [25] and [48] 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%. Surprisingly, we also improve upon our best CelebA-C+A results. This is partially

Method	Accuracy
MOON [24]	90.94%
CLMLE [29]	91.13%
MCNN-AUX [26]	91.29%
AFFACT [25]	91.67%
SSP + SSG [27]	91.80%
FAN [28]	91.81%
HFE [31]	92.17%
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 3.1: 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
LNets+ANet [10]	87%
Zhong et. al. [47]	89.80%
AFFACT [25]	91.45%
AFFAIR [48]	91.45%
SA [6]	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%

Table 3.2: Comparison between our baseline ResNet-18 networks and state-of-the-art methods on the in the wild images (CelebA-ITW).

because the data was labeled using the original images, and some attributes are not visible in the aligned version. In particular, *wearing necklace* is 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. For example, the *attractive* attribute may be affected by the clothing worn in the image, which is mostly cropped out by alignment.

Perhaps even more surprisingly, we are also able to obtain results comparable to state-of-the-art with just 10% of the training data available in CelebA (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, including [25] [27] [32]. All methods which improve upon our CelebA-ITW results, with or without ImageNet pretraining, use either a pretrained face detector or additional data.

3.1.2 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.

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

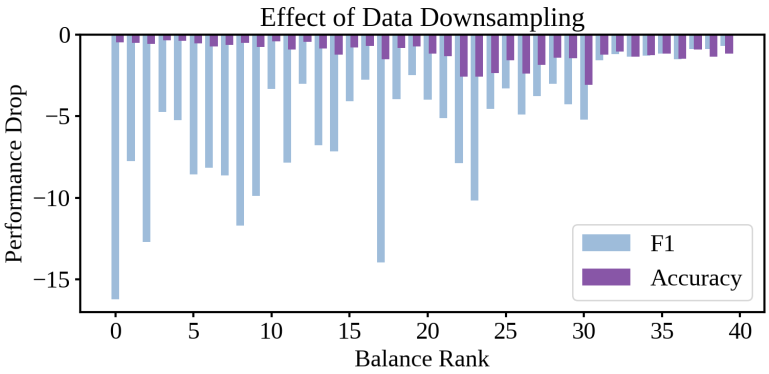


Figure 3.1: Per-attribute accuracy and F1 drop incurred by training on a random selection of 10% of the training data. Balance rank orders attributes by their ratio between positive and negative samples, with rare attributes (e.g. *bald*) on the left and common attributes (e.g. *young*) on the right.

recall (combined using F1), as shown in Figure 3.1. Because there is little improvement than can be made over always predicting the majority class, performing well for highly imbalanced attributes is not very important for achieving high average accuracy scores.

To address this issue, several previous works [27–29, 49, 50] have used balanced accuracy, which weighs true positive rate equally to true negative rate:

$$\frac{1}{2}(\frac{TP}{TP+FN}+\frac{TN}{TN+FP}) \tag{3.1}$$

Average precision has also been used [27]. 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 [51]. 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

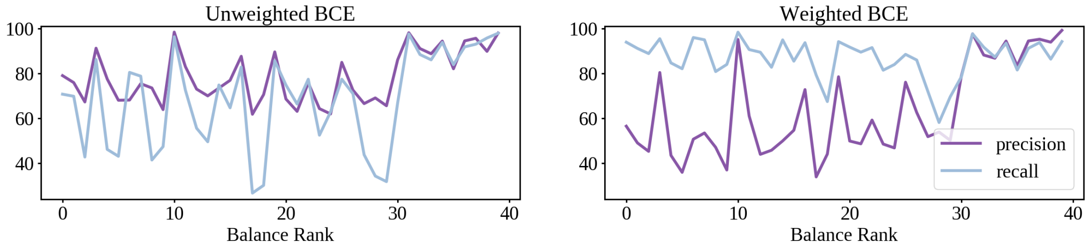


Figure 3.2: Precision and recall for a ResNet-18 network optimized with BCE (left) and balance-weighted BCE (right). Balance rank orders attributes by their ratio between positive and negative samples, with rare attributes (e.g. *bald*) on the left and common attributes (e.g. *young*) on the right.

by rewriting balanced accuracy as the following:

$$\frac{1}{2}(1-\frac{FN}{N_p})(1-\frac{FP}{N_n}), \tag{3.2}$$

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 3.2. 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 at-

	Without Pretraining		
Data	Acc.	Bal. Acc.	F1
CelebA-C+A, 100%	91.48 ± .06%	81.49 ± .18%	71.98 ± .21%
CelebA-C+A, 10%	90.32 ± .07%	78.20 ± .43%	66.58 ± .76%
CelebA-ITW, 100%	91.36 ± .13%	82.80 ± .11%	73.13 ± .21%
CelebA-ITW, 10%	89.86 ± .13%	76.34 ± .72%	63.72 ± 1.58%

	With ImageNet Pretraining		
Data	Acc.	Bal. Acc.	F1
CelebA-C+A, 100%	91.71 ± .01%	82.35 ± .11%	73.19 ± .10%
CelebA-C+A, 10%	90.88 ± .02%	79.62 ± .12%	69.05 ± .11%
CelebA-ITW, 100%	91.81 ± .07%	82.39 ± .25%	73.36 ± .31%
CelebA-ITW, 10%	90.43 ± .05%	76.92 ± .19%	64.57 ± .28%

Table 3.3: Accuracy, balanced accuracy, and F1 scores averaged over all attributes.

tribute 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 ± .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 ± .13% to 58.62 ± .10% and our accuracy drops from 91.72 ± .01% to 86.10 ± .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, [29] compares their model to [27] 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 [52] 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 3.3. 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 [28], 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.

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 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 × 178 to 274 × 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 3.3. 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. [6] 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, [28], 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.

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 [32]. 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 by averaging the 200 validation images which result in the highest activations for each attribute. For example, as shown in Figure 3.3, *narrow eyes* frequently applies to partially closed eyes due to laughing and *high cheekbones* seems to just recognize smiling (we find that

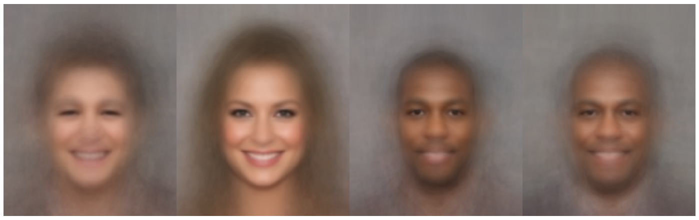


Figure 3.3: Average of the 200 validation images which achieve the highest activations for *narrow eyes*, *high cheekbones*, *big lips*, and *big nose* (left to right).

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 3.4, *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

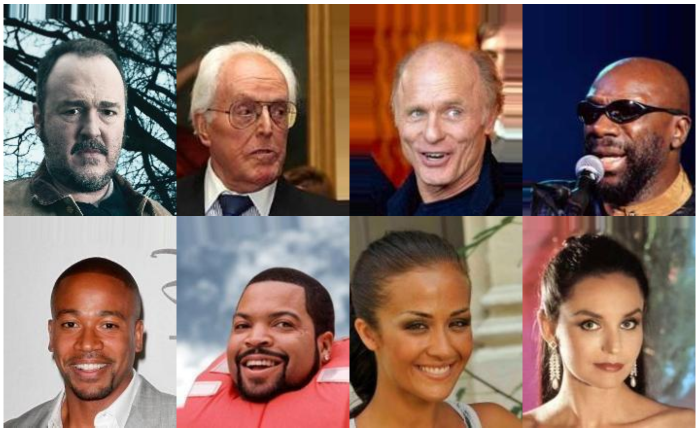


Figure 3.4: Top row: Examples of validation images labeled as *bald*. Some images are clearly not bald (leftmost example) or clearly bald (rightmost example), but there is some ambiguity in between. Bottom row: examples of validation images labeled as *receding hairline* seemingly due to close-cropped or tied back hair.

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, 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 [32] 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 3.4.

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

Table 3.4: 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.

3.1.3 Summary

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.

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.

3.2 Data Issues

Despite the popularity of CelebA, there are a multitude of widespread, unaddressed attribute labeling issues. While the subjectivity of many attributes in the dataset makes complete analysis difficult, the majority of labels we are able to analyze have a large number of errors or inconsistencies. We use several techniques to evaluate label quality. We first create a list of contradicting attributes and find that 6.78% of images are labeled with attributes which directly contradict one another. We then relabel a random sample of 400 images for all non-subjective attributes and find that some attributes have false positive rates as high as 25%, while others have false negative rates as high as 22%. To evaluate subjective attributes, we use age estimation and semantic segmentation to provide estimates of age and feature size, and compare these estimates with the binary attributes in CelebA. We find that such attributes are highly inconsistent with these more fine-grained measures, preventing even near-state-of-the-art classifiers from achieving reasonable performance. Finally, we show that some attributes are correlated in ways that cannot be explained by dataset imbalance, indicating incorrect labeling or gender and racial bias. In total, we determine that at least 10 of the 40 attributes in CelebA have major issues such as frequent contradictions, incorrect labels, or significant inconsistency.

3.2.1 Incorrect Labels

We first focus on labels which can be directly shown to be incorrect. For subjective labels, we do this by identifying contradicting labels. For non-subjective labels, we manually relabel random samples to determine the frequency of incorrect labels.

Label	Contradictions	%
No Beard	5 o'Clock Shadow, Goatee, Mustache	4.0%
5 o'Clock Shadow	Goatee, Mustache, No Beard	47.9%
Straight Hair	Wavy Hair	2.7%
Bald	Bangs, Receding Hairline, Straight Hair, Wavy Hair	33.3%

Table 3.5: Contradicting attribute labels. % is the percentage of images in the full dataset with the label in the left column and a contradicting label from the middle column.

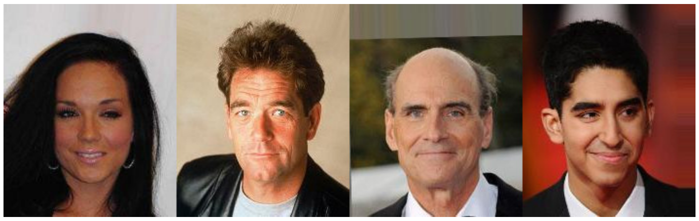


Figure 3.5: The first four images in CelebA labeled as *double chin* but not *chubby*. None of these images contain a double chin.

Contradicting and Conflicting Labels

To determine the prevalence of incorrect labels for subjective attributes, we first count the number of labels which are contradicting (in direct opposition to one another). For example, it is not possible to have both *straight hair* and *wavy hair*. To determine how many labels in CelebA directly contradict another label, we define a list of all contradicting attributes. This is shown in Table 3.5. *No beard* contradicts with all facial hair labels other than *sideburns*. Depending on definition, *sideburns* may also contradict with *no beard*, but we find that this only applies to 128 images (0.06% of the dataset) so we do not include it. Similarly, *5 o'clock shadow* contradicts with other facial hair, *straight hair* contradicts with *wavy hair*, and *bald* contradicts with all hair labels. We find that 6.78% of images have at least one contradicting label based on this list, and that *bald* and *5 o'clock shadow* contradict with another label in one third or more images labeled with these attributes.

We also find that there are many labels, which, while not necessarily contradicting, conflict with one another. 2.33% of images labeled with a hair color are labeled with multiple hair colors, most commonly either both *brown hair* and *black hair* or *brown hair* and *blond hair* due to the unclear separation between classes. We also find that 38.1% of images labeled with *double chin* are not labeled with *chubby*. While *double chin* does not necessitate *chubby*, this is frequently indicative of bad labeling, as shown in Figure 3.5. Similarly, while hair color labels don't contradict with *bald* because they may refer to facial hair, we find that this is very rarely the case. If we add these conflicts to our list of contradictions, we find that 9.84% of the dataset contains at least one pair of contradicting labels, and the contradiction frequency for bald rises to 42.3%.

Mislabeling

To determine the prevalence of incorrect labels for non-subjective attributes, for each attribute we construct one randomly sampled subset of 400 images containing only positive instances, then another containing only negative instances. We then manually verify the correctness of the labels. To avoid sampling bias, a random seed of 0 is used for all attribute samples. Results are shown in Table 3.6. We find that there are very few entirely non-subjective labels in CelebA; of the 40 total attributes, only 7 can be clearly defined. Even for these labels, there is some ambiguity. For *wearing hat*, we assume that hoods and bandanas count as hats. Without this assumption, the number of false positives rise to 26 and false negatives drop to 5. We also find that *mouth slightly open* is better defined as *mouth open*. Images of people with wide open mouths are consistently labeled as *mouth slightly open*, but images of people with slightly open mouths are labeled inconsistently. Examples of incorrectly labeled images are shown in Figure 3.6. In addition to incorrect labels, we find that in 86 images (21.5%) correctly labeled as *wearing necklace*, the necklace is entirely cropped

Label	**FP**	**FN**
Eyeglasses	2 (0.5%)	0 (0.0%)
Mouth Slightly Open	5 (1.3%)	86 (21.5%)
Male	5 (1.3%)	2 (0.5%)
Wearing Hat	14 (3.5%)	9 (2.3%)
Wearing Earrings	53 (13.3%)	44 (11.0%)
Wearing Necklace	102 (25.5%)	39 (9.8%)
Wearing Necktie	56 (14.0%)	3 (0.8%)

Table 3.6: Number of false positives (FP) and false negatives (FN) for non-subjective attributes out of a sample of 400. Gender labels were verified using identity labels.

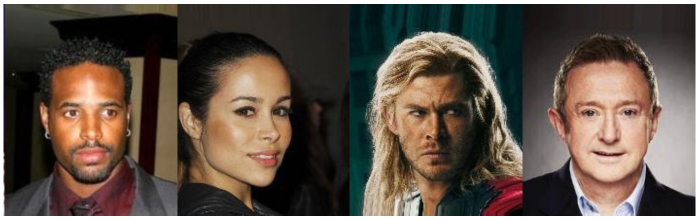


Figure 3.6: Examples of false positives (left to right): *eyeglasses*, *wearing earrings*, *wearing hat*, and *wearing necktie*.

out in the aligned version. In many more images, the necklace is visible but too small or similar to clothing to be noticeable. This makes accurately predicting *wearing necklace* near-impossible for the aligned version of the dataset.

Using the false positive and false negative rates in Table 3.6 combined with attribute probabilities and correlations, we estimate that 34.3% of images in CelebA have at least one incorrect label among these seven. Note that almost all attributes are predominately negative (77% of all labels are negative), so the contribution of false negatives is far greater than the contribution of false positives. Importantly, incorrect labels cannot be treated as random noise. Of the 102 images incorrectly labeled with *wearing necklace*, 100 are of women. Of the 56 images incorrectly labeled with *wearing necktie*, all are of men, most of whom are wearing a collared shirt and coat as shown in Figure 3.6. We therefore suggest that these labeling issues were likely caused by labelers misunderstanding a set of reference images, resulting in systemic mislabeling. Details about CelebA data collection are not provided, so we are unable to determine the specific cause of these issues. These errors are far more problematic than random noise because classifiers are able to learn the noise. For example, a classifier trained on CelebA may predict that someone wearing a collared shirt is wearing a necktie even if they are not, because this is frequently the case in the training data.

3.2.2 Inconsistent Labels

While incorrect labels are an issue for many attributes, most attributes are subjective and therefore cannot be directly relabeled or shown to contradict with other attributes. We instead show that many subjective attributes fail to capture quantitative information about the feature they describe or are strongly correlated with other, unrelated attributes.

Consistency

To evaluate label quality in subjective attributes, we take advantage of other facial analysis tasks that can be used to estimate quantitative information about subjective CelebA attributes. Semantic segmentation can be used to estimate the size of different facial regions, and age estimation can be used to estimate youth. We therefore compare all attributes which subjectively label the size of facial features (*big lips*, *big nose*, and *narrow eyes*) as well as *young*, which subjectively labels the age of the face, with these classifiers. We find that the subjective labels are highly inconsistent with respect to these quantitative metrics, preventing even near-state-of-the-art classifiers from achieving acceptable performance.

For age estimation, we use DEX [53] to estimate the age of all images in CelebA. For semantic segmentation, we use the DeepLabv3+ architecture [54] trained on the

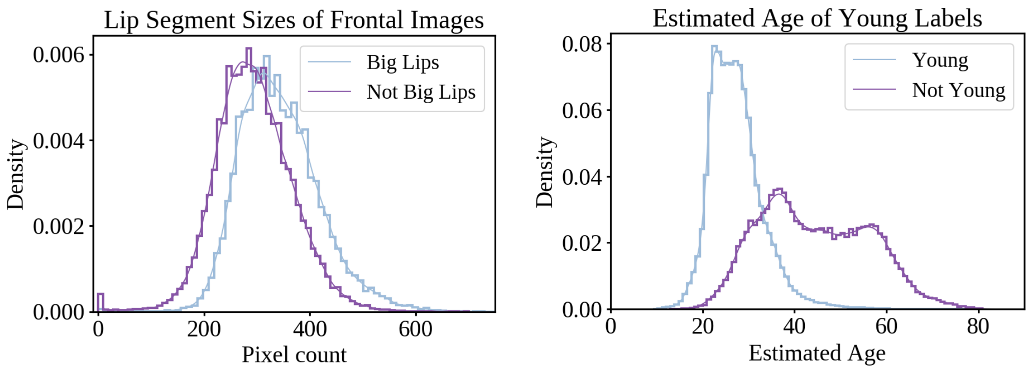


Figure 3.7: Left: Histogram of combined lip segment size for images labeled with *big lips* and images not labeled with *big lips*. Right: Histogram of estimated ages for images labeled with *young* and images not labeled with *young*.

CelebA-Mask-HQ dataset, which annotates 18 facial regions for the $30,000$ image CelebA-HQ dataset [18]. We predict part masks for all images in CelebA. Because segment size is affected by pose, we restrict our analysis to frontal images to ensure consistent evaluation of part size. We use the HopeNet pose estimation network [55] to estimate head poses for all images in CelebA and discard all images with a pitch or roll not within $\pm 10°$ or a yaw not within $-20°$ and $5°$. Because CelebA images are generally frontal, this still leaves $84,970$ out of $202,599$ images for analysis.

To evaluate the consistency of size-based attributes we count the number of pixels contained in the segment associated with each attribute. We find that the attribute labels provided by CelebA do a poor job of discriminating these features. Images labeled with *big lips* have an average lip size of 343.1 ± 75.4, and images not labeled with *big lips* have an average lip size of 293.5 ± 73.2. This is shown in Figure 3.7. We obtain similar results for *big nose* and *narrow eyes*: images labeled with *big nose* have an average nose size of 560.97 ± 69.81, with all other images having an average nose size of 518.96 ± 68.42. Images labeled with *narrow eyes* have an average eye size of 72.61 ± 33.42, with all other images having an average eye size of 104.01 ± 33.07. A linear classifier trained to predict these attributes using segment sizes (assuming a

Attribute	Estimated	ResNet-18
Narrow Eyes	38.31	45.47
Big Lips	52.47	46.73
Big Nose	46.34	64.89
Young	90.05	92.91
Mouth Slightly Open	93.93	95.70
Eyeglasses	95.03	98.10

Table 3.7: F1 scores for linear classifiers using estimated quantitative metrics to predict subjective labels compared with a ResNet-18 classifier trained on CelebA. All attributes other than *young* are evaluated using only frontal images.

balanced distribution) is unable to reach an F1-score on the test data above 50 except for when predicting *big lips*, for which it achieves an F1-score of 52. This indicates the actual size of the features has little bearing on whether labelers described them as "big." As shown in Figure 3.7, we find that the *young* attribute is far more consistent, but still cannot be predicted completely reliably. The substantial overlap in estimated age between positive and negative instances demonstrates that even with reasonably consistent labeling, subjective binary attributes are highly flawed for representing non-binary features. A model with higher accuracy for *young* than a competitor may simply do a better job of capturing labeling bias than actually estimating age.

Even for less inconsistently labeled attributes, we find a large amount of overlap between positive and negative samples. We use DEX [53] to estimate the age of all images in CelebA, then compare this with the *young* attribute. While there is a far more clear separation between the distributions than for the size-based attributes, we still find there is a great deal of overlap. A classifier trained to predict *young* based on estimated age can only achieve an F1-score of 90, by predicting all people with an estimated age below 33.61 years old to be young.

To demonstrate that the poor performance of these classifiers is not a result of bad segmentation and age estimation, we provide comparison to our pretrained ResNet-18 classifier described in Section 3.1. These results are shown in Table 3.7. This classifier does not perform substantially better than using segment size or age estimation, and even performs worse for *big lips*. To demonstrate that non-subjective attributes can be accurately estimated using quantitative classifiers, we also estimate *mouth slightly open* and *eyeglasses* using the mouth and glasses segments. While *mouth slightly open* is not entirely consistently labeled (as discussed in section 3.2.1) and glasses propped on foreheads cause issues for our segment-size based classifier, we are still able to achieve satisfactory performance. We therefore suggest that subjective, size-based labels are too inconsistent for any classifier to achieve reasonable performance. Other highly subjective labels likely suffer from similar issues.

To further show the extent to which these labels are inconsistent, we compute the agreement across labels for different images of the same person. We compute the Fleiss' κ metric for each attribute across each person, where labels for different images are treated as different raters. We then compute the average κ across all subjects to determine the average agreement. These agreements are shown in Table 3.8. Note that κ ranges from -1 to 1, where $\kappa < 0$ indicates even less agreement than expected by chance and $\kappa > 0$ indicates some amount of agreement, up to perfect agreement at $\kappa = 1$. As shown in the table, only 12 attributes have agreement higher than $\kappa = .5$.

Correlated Labels

Counting contradictions, relabeling, and evaluating consistency with a quantitative classifier still leaves many attributes unanalyzed. While we are unable to directly evaluate the quality of these attributes, there are some correlations between subjective attributes which indicate poor labeling. As discussed by previous work, on average attributes have a gender skew of 80.0% [34]. For example, 27.9% of images labeled with *male* are labeled with *attractive*, whereas 67.9% of images not labeled with *male* are labeled with *attractive*. It is difficult to tell to what extent this is a result of bias in

Blurry	−0.0181	Wavy Hair	0.4272
Pale Skin	0.1562	Bangs	0.4313
Mouth Slightly Open	0.2141	Pointy Nose	0.4546
Narrow Eyes	0.2378	Black Hair	0.4717
Wearing Hat	0.2489	Sideburns	0.4759
Smiling	0.2551	Bushy Eyebrows	0.4893
Wearing Necktie	0.2712	Mustache	0.4945
Double Chin	0.2910	Bald	0.4981
Wearing Necklace	0.3113	Goatee	0.5062
High Cheekbones	0.3178	5 ’o Clock Shadow	0.5131
Rosy Cheeks	0.3282	Arched Eyebrows	0.5131
Bags Under Eyes	0.3388	Attractive	0.5140
Receding Hairline	0.3405	Big Nose	0.5585
Straight Hair	0.3441	Blond Hair	0.5727
Brown Hair	0.3719	Heavy Makeup	0.6302
Oval Face	0.3881	No Beard	0.6450
Wearing Earrings	0.3893	Big Lips	0.7279
Chubby	0.3924	Wearing Lipstick	0.7322
Eyeglasses	0.4150	Young	0.8360
Gray Hair	0.4233	Male	0.9789

Table 3.8: Average Fleiss κ agreement for the 40 attributes in CelebA

labeling rather than bias in data selection, but there are some cases were correlation is clearly indicative of bad labeling. The clearest example is *high cheekbones*, which has a correlation of 0.68 with *smiling*. 85.6% of images labeled *high cheekbones* are also labeled *smiling*, which is otherwise only applied to 48.2% of all images. This is likely because cheekbones appear higher while smiling, particularly when the smile is wide. Therefore, it is highly unlikely that *high cheekbones* provides an accurate label of cheekbone height irrespective of expression. Additionally, some gender correlations are too strong to be explained by data selection. We find that women are 3.1 times more likely to be labeled with *pointy nose*, whereas men are 2.9 times more likely to be labeled with *big nose*. This is despite the fact that the probability of a random male nose being larger than a random female nose in terms of segment size is just 54.3%, indicating gender bias substantially influences labeling.

Gender bias is also not the only bias encoded by subjective attributes. While the correlation between *big lips* and *big nose* in the validation set is fairly weak (0.054), our ResNet classifier described in Section 3.2.2 exaggerates this correlation to 0.091 due to related biases. Analysis of the 200 images which achieve the highest activations for these attributes show that they are heavily biased towards black men. 99% of people in the top activations for *big lips* are black, and 94% are male. 78% of people in the top activations for *big nose* are black, and 97% are male. None of the people in the images with the 200 lowest activations for either attribute are black. The bias exhibited for *attractive* is also exaggerated, with the percentage of men predicted to be attractive dropping to 25.1% while the percentage of women rises to 71.0%.

3.2.3 Summary

While the subjectivity of CelebA labels makes their quality difficult to evaluate, we find that most labels we are able to quantitatively evaluate are poorly or inconsistently used. In particular, 10% or more instances of *5 o'clock shadow*, *bald*, *wearing earrings*, *wearing necklace*, and *wearing necktie* are used incorrectly or contradict another label. *Mouth slightly open* is labeled consistently enough to predict reliably, but the predictor does not match the label definition. Furthermore, subjective labels such as *big nose*, *big lips*, *narrow eyes*, and *young* are inconsistent with a quantitative classifier measuring the same feature. Attributes can also be shown to be poorly labeled through correlations. *High cheekbones* almost entirely overlaps with *smiling* and *pointy nose* is strongly negatively correlated with *male*. Other attributes clearly encode bias which is amplified by a classifier trained to predict those attributes. *Big lips* and *big nose*, while doing a poor job of estimating quantitative measures of lip and nose sizes, both encode racial bias which is learned by a classifier. In total, we find that there are 5 attributes which are clearly labeled incorrectly or contradict with another attribute more than 10% of the time, and another 5 attributes which are highly inconsistent or can be shown to be highly problematic through correlation with another attribute. There are many other attributes we are unable to evaluate, but are likely also poorly labeled. For example, a surprisingly high number (34.8%) of images labeled *male* are also labeled *bags under eyes*, and our ResNet classifier is unable to achieve an F1 greater than 50 for *oval face*.

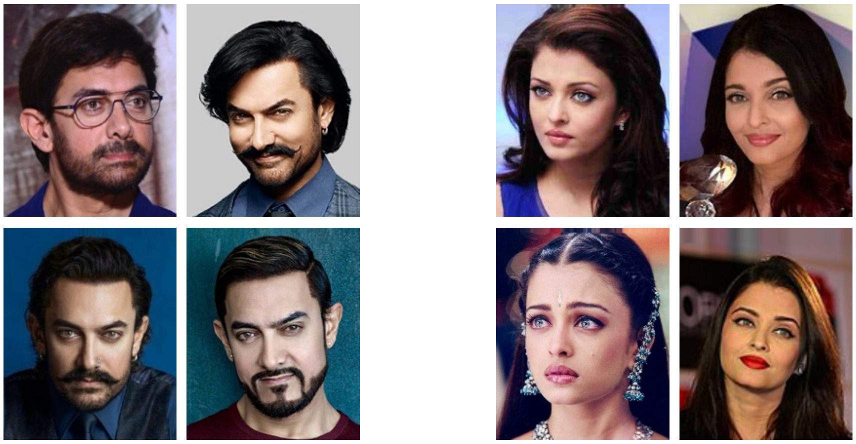
Chapter 4

Utilizing Attributes for Face Captioning

In this chapter we investigate the applicability of attributes for natural language face description. Section 4.1 deals with the use of weakly labeled features to overcome the lack of high-quality data available in CelebA, and section 4.2 deals with the generation and evaluation of face captions using CelebA attributes and our proposed weak labels.

4.1 Weakly Labeled Prominent Features

In this section we propose a method for generating weak prominent feature labels to help overcome issues with existing labeled data. We collect a small dataset of natural language descriptions for 205 people and construct a classifier which uses semantic segmentation to predict these features. We show that we are able to accurately represent 12 segment-based features which we use to generate 22 distinct binary prominent feature labels.



Labels: well-defined nose tip, ears stick out, pierced ears, arched eyebrows, puffy eyelids, bags under eyes, mustache, goatee.

Labels: wide eyes, long eyelashes, hooked nose, thin nose, small head, full lips, arched eyebrows.

Figure 4.1: Examples of two subjects in the caricature set and their associated ground-truth labels.

4.1.1 Data

We first collect a dataset consisting of 2,900 images of 205 celebrities. All images are manually cropped to 218x178 boxes containing the face and hair and manually labeled with at least four prominent features such as “wide nose,” “high cheekbones,” and “pointy chin” by two graduate students and a professor at the University of Nevada, Reno. Two labelers independently annotated all 205 celebrities, and the third acted as a tie-breaker in cases of disagreement. The only restriction we imposed on the prominent feature selections was that each should consist of a part and a description (e.g. “nose,” and “wide”). In total, the dataset contains an average of 7.04 features per person and a total of 150 unique features; our weak labeling method described in the following section covers 36 of these features. The dataset also contains 1,424 caricature images which are not used for this work. We refer to the dataset as the “caricature dataset.” Two examples are shown in Figure 4.1.

4.1.2 Weak Labeling

To generate weak labels, we first use semantic segmentation to determine the locations and sizes of each region in the face. Face images are segmented into 18 different classes based on the labels provided by the CelebAMask-HQ dataset [18]. These classes include face, nose, eyes (left/right), lips (upper/lower), mouth, ears (left/right), eyebrows (left/right), hair, neck, and various items (glasses, hats, clothing, earrings, and necklaces). Segments are allowed to overlap; for example, there may be hair partially covering the forehead or ears. In the CelebAMask-HQ dataset, labelers attempted to segment partially occluded regions but did not segment mostly or fully occluded regions (for example, if an ear is completely covered by hair it is not segmented).

For our segmentation model we use DeepLab v3+ [54] with a ResNet [46] backbone pretrained on the ImageNet dataset [56]. After pretraining, the full model is trained to segment all 18 facial features labeled in the CelebAMask-HQ dataset. To improve generalization, the dataset is augmented with random rotation, grayscaling, blurring, and elastic deformation [57]. We also downscale CelebAMask-HQ by one half to 512 × 512 and apply random scale augmentation with a modifier between .5x and 1.25x. Because DeepLab does not work very well with images at a resolution much lower than this, during inference we rescale our images from 218 × 178 to 418 × 512. Figure 4.2 shows an example of a mask output by the network for an image in the caricature dataset.

20 shape measurements are extracted from the segmentation results and used as features to predict 36 shape-based prominent facial feature labels. For each subject, non-frontal images are discarded and the remaining images are rotated to be horizontally level. Most features are based on basic distance measurement. For example, eye distance is defined as the distance between eye centroids, nose height is defined as

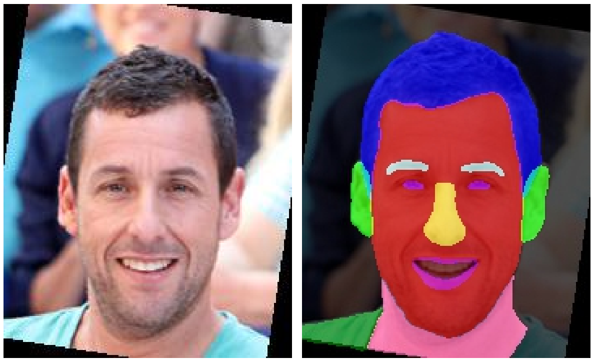


Figure 4.2: Visualization of a segmentation mask output by the DeepLab v3+ network.

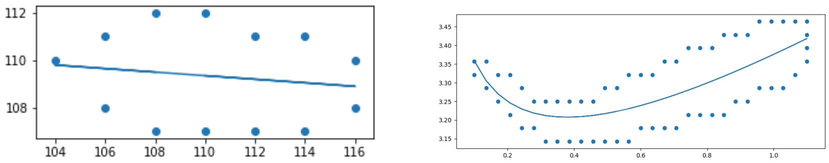


Figure 4.3: Plots of an $ax + b$ function fit eye contours (left) and an $alog(x) + bx + c$ function fit to eyebrow contours (right).

the distance between the highest and lowest nose point, and lip thickness is defined as lip area divided by mouth width. There are also two ratio-based features: face aspect ratio (the ratio between face width and face height) and face size (the ratio between face area and head area). Finally, there are some features calculated using the parameters of a function fit to the shape contours. Eyebrow arch is defined as the b parameter of a $alog(x) + bx + c$ function fit to the eyebrow, chin flatness is defined as the p parameter of a x^p function fit to the bottom of the head, and eye angle is defined as the slope of a linear function fit to the eye. Complete details are provided in Table 4.1. To demonstrate why we use a log function for eyebrows and a linear function for eyes, we plot examples of functions fit to these regions in Figure 4.3.

Feature	**Definition**
Ear size	Area of ear segment (averaged between left/right ear)
Ear width	Width of smallest bounding box containing the ear (averaged between left/right ear)
Ear height	Height of smallest bounding box containing the ear (averaged between left/right ear)
Eye distance	Distance between eye centroids. Normalized by width of head segment.
Eye size	Area of eye segment (averaged between left/right eye)
Eye angle	a parameter of $ax + b$ function least-squares fit to eye (averaged between left/right eye)
Eyebrow arch	b parameter of $alog(x) + bx + c$ function least-squares fit to eyebrow (averaged between left/right eyebrow)
Eyebrow width	Arc length of the log function described above (averaged between left/right eyebrow)
Eyebrow height	Area of eyebrow segment (averaged between left/right eyebrow). Normalized by eyebrow width.
Chin flatness	p parameter of x^p function least-squares fit to region from bottom of head to halfway between bottom of head and lower lip. This region is normalized to have a width of 2 and a height of 1 and is centered at the origin.
Face aspect ratio	Aspect ratio of smallest bounding box containing the face. The face is defined as the eyes, nose, and upper lip (lower lip not included to reduce error resulting from smiling)
Face size	Area of smallest bounding box containing the head (entire face segment predicted by the network) divided by area of smallest bounding box containing the face (defined above)
Forehead size	Area of smallest bounding box containing the head, cut off below the eyes.
Glasses	Binary flag: 1 if the glasses segment has nonzero area, 0 otherwise.
Lip thickness	Area of lip segment (averaged between upper/lower lip). Normalized by mouth width.
Lip thickness ratio	Area of upper lip segment divided by area of lower lip segment
Mouth width	Width of smallest bounding box containing the upper and lower lip
Nose width	Width of smallest bounding box containing the nose
Nose height	Height of smallest bounding box containing the nose

Table 4.1: Shape measurement computation details. All metrics are normalized by distance between the eyes unless otherwise specified

4.1.3 Evaluation

To evaluate the similarity between the segmentation-based shape features and what humans perceive to be prominent features, we compute how correctly the generated weak labels order the ground truth descriptions in the caricature dataset. Each of the 20 shape features is associated with either one or two of the 36 labels. For example, the "nose width" shape feature is associated with the "wide nose" and "thin nose" labels. To determine the accuracy of a shape feature, all 203 are subjects are ordered by that shape feature. Because we have several images of each subject from which the feature can be measured, we average each measurement across all images of a subject. To avoid noisy estimations for subjects with few images, we construct a prior distribution for each feature using the entire dataset then compute the posterior distribution for a particular subject as

$$\mu_{post} = (\frac{n/\hat{\sigma}^2}{1/s^2 + n/\hat{\sigma}^2})m + (\frac{1/s^2}{1/s^2 + n/\hat{\sigma}^2})\hat{\mu} \qquad \sigma^2_{post} = (\frac{1}{s^2} + \frac{n}{\hat{\sigma}^2})^{-1},$$

where n is the number of images of the subject, $\hat{\mu}$ and $\hat{\sigma}$ are the sample mean and sample variance of the feature for that subject, and m and s^2 are the prior mean and variance of the feature. The mean of the posterior is used for ranking subjects by the feature.

To determine the quality of each feature, we compute how useful that feature is for predicting its associated label(s). This is done by calculating the optimal information gain for the labels from splitting the subjects on that feature. Information gain is a common metric for determining the usefulness of a feature for predicting a label, and has long been used for the induction algorithm in decision trees [58]. For features with multiple labels, we split subjects into three categories (negative label, no label, positive label) based on the estimated value of the feature. For features with one

Feature	**Predicted labels**	**Entropy**	**Gain**	**Ratio**
eye distance	wide-set, narrow-set	1.151	.166	.144
nose width	wide, wide nostrils, thin	1.150	.348	.303
nose height	long, short	.483	.142	.294
brow height	bushy, thin	.887	.364	.410
lip thickness	pouty/full, thin	1.216	.439	.362
chin flatness	square, pointy	.981	.0920	.0938
eye size	wide, narrow	1.418	.399	.281
mouth width	big, small	1.221	.257	.211
ear size	big, small	.828	.0984	.119
forehead size	big, small	.973	.188	.193
face aspect ratio	long, wide	.478	.138	.289
face size	small, big	.682	.0794	.116
ear width	stick out, flat	.955	.296	.310
brow coeff	arched	.718	.177	.246
ear height	high, low	.351	.0710	.202
glasses	glasses	.263	.174	.661
brow width	long, short	.853	.175	.205
eye angle	slanted up, slanted down	1.25	.277	.221
lip thickness ratio	thin upper, thick lower	.791	.0630	.0797

Table 4.2: Entropy, information gain, and the ratio between the two for each feature/label combination in the caricature dataset.

label, we split subjects into two categories (positive label, no label). We then compute information gain as $G(x, f) = H(x) - H(x|f)$, where $H(x|f)$ is the conditional entropy of the data given cutoffs for feature f. For example, the labels for "arched eyebrows" have an entropy of .718. Splitting the subjects into "brow coefficient $<$ .477" and "brow coefficient $>$ .477" results in a conditional entropy of .541, so the information gain is .177. Because information gain is impacted by the entropy of the feature (e.g., if the label already has low entropy it is hard to gain any information), we also provide the ratio between information gain and original label entropy. These results are shown in Table 4.2.

Note that even for seemingly easy to predict features, perfectly ordering the subjects is difficult. "Glasses," for example, is reasonably easy to predict as the segmentation of this feature is generally very accurate. However, there is not a perfect correspondence between how frequently at which a subject wears glasses and how likely labelers were to consider the glasses a prominent feature, so the feature is only able to reduce entropy from .263 to .174 (a ratio of .661).

4.1.4 Attribute Prediction

We use the described weak labeling method to predict prominent feature labels for all images in CelebA. To ensure we are only using high-quality weak labels, we only utilize features which achieve an information gain to entropy ratio greater than 0.2 and do not already exist in CelebA for our captions described in the following section. These labels are then used to generate captions as described in the next section. The primary drawback to the approach described in Section 4.1.2 is that we require images to be entirely frontal. However, most images are not perfectly frontal so this drastically limits the system. To deal with this, we use a face frontalization network to rotate images before generating labels. This is not a perfect process due to the inherent ambiguities present in rotated images, but allows us to make reasonable predictions for most images. We use Rotate-and-Render [59] for face frontalization. This is shown in Figure 4.4. Rotate-and-Render fails for many more difficult images, so we create an alternative set of weak label which rely on ground-truth identity labels. Features which should remain consistent across different images of the same person (e.g., big nose), are averaged across all images of a person to reduce noise from poor segmentation or frontalization. In the following section we provide results with both the raw, per-image labels and the identity-averaged labels.

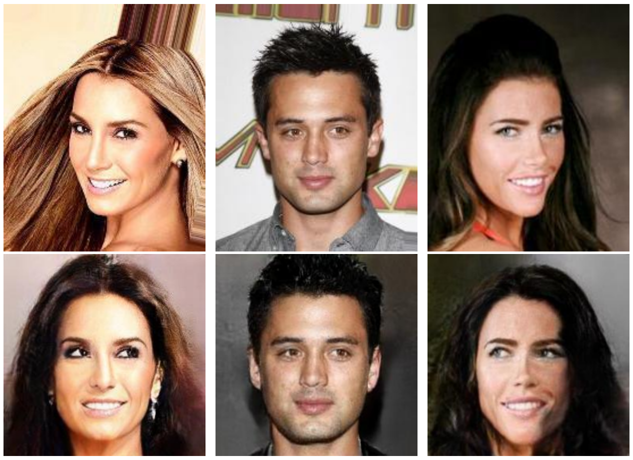


Figure 4.4: Examples of images at different orientations rotated using Rotate-and-Render

4.2 Captioning

In this section we propose the problem of face captioning and discuss methods for generating and evaluating captions. Face captioning is useful for the following applications:

- Accessibility. Current captioning methods are not able to describe what a face looks like in a way that would allow for someone to build a mental image of the person.

- Lookup. Captioning allows for image lookup based on text similarity. It is far more simple to describe a face than to draw it, but current image lookup methods for faces are based entirely on visual similarity. This means that someone would have to draw the face in order to find it, which many people are unable to do.

- Image Manipulation. It has recently become popular to perform image manipulation using text captions, as this is easier for humans to control than prior image manipulation techniques using segmentation.

Despite these applications, to our knowledge there is no prior work investigating how to best generate captions for face images and evaluate the quality of generated captions. Unlike simpler problems such as describing the color and general shape of a bird, describing human faces is highly subjective and detail-oriented. We therefore propose a simple but useful method of generating face captions and two techniques for evaluating the quality of these captions.

4.2.1 Implementation

We first describe our method for generating image captions from attribute labels. Prior works have used a PCFG to generate captions using CelebA attribute labels [11, 12]. For our baseline we use a similar technique, but rather than grouping all attributes into "wearing," "has," and "is" we group attributes into ones that can function as nouns and ones that can function as adjectives (many can function as both). Nouns are then split into description of qualities of the face (e.g., hair and feature size), description of expression (smiling, mouth open), descriptions of clothing or makeup being worn (wearing hat, wearing lipstick), and descriptions of the image itself (blurry). Qualities which apply to the same feature (e.g., big nose and pointy nose) are grouped together to avoid awkward sentences which refer to the same part multiple times. This allows us to construct more grammatical sentences that align more closely with how an actual person would write face descriptions. To avoid overly verbose sentences, we split captions into multiple sentences which then can be reordered for data augmentation. Comparison between captions generated by our method and those provided by Multi-Modal CelebA sample-captions are shown in Table 4.3. For fairness of comparison, we randomly discard a subset of attributes to keep captions concise as is done by prior works. For the provided captions, each attribute has a 33% chance of being discarded.

The person has high cheekbones, and pointy nose. She is wearing lipstick.	A woman with bags under her eyes, big lips, long eyebrows, and brown hair. Her mouth is wide open and she is smiling.
This person has arched eyebrows, wavy hair, and mouth slightly open. She wears lipstick. She is attractive.	A woman with blond hair and big lips. Her mouth is open.
She wears lipstick. She is smiling, and attractive and has wavy hair, and brown hair.	An attractive woman with a long nose, brown hair, and high cheekbones.
This woman wears lipstick. She has bushy eyebrows, and big nose. She is young, and smiling.	A picture of an attractive woman with black hair. Her mouth is open and she is smiling.
She has mouth slightly open, straight hair, and big lips. She wears earrings. She is young.	An image of a woman with wide-set eyes and straight hair. Her mouth is slightly open. She is wearing earrings.

Table 4.3: Comparison between randomly-generated captions in Multi-Modal CelebA (left) and captions in our datast (right) for the first five images in CelebA-HQ.

Because CelebA captions are highly inconsistent across different images of the same person, we construct an alternate version of CelebA in which attributes which are not image specific are forced to be consistent for all images of a single person (things such as big nose, oval face, and high cheekbones should not vary between different images of the same person). The consistent label assigned to each person is the rounded average (i.e., majority vote) of the labels for all images of that person. This helps preserve features that are prominent for a particular person between different images.

Because in practice ground-truth attributes will not be available to generate captions, we train classifiers to predict CelebA attribute labels and shape-based labels and use these to generate the captions used for evaluation. For CelebA attributes, we use the pretrained ResNet-18 model described in Section 3.1.1. For predicting the weakly labeled shape-based labels, we use an identical model trained to predict the percentile of each shape feature. We use mean-squared error as our loss function and train for 20 epochs using the Adam optimizer.

4.2.2 Evaluation

We suggest two important qualities in an evaluation metric based on the previously described use cases. First, captions should describe the specific prominent features of the person in the image, such that a person or a model can predict with high accuracy if two captions are of the same person. This evaluates the extent to which the description captures unique features which someone could use to recognize the person being described in a different context in which features such as clothing and makeup may change. This metric is the main focus of our weak label-based captions. Second, captions should describe the features present in the specific image, such that someone describing an image of a person can easily find the correct image in a large group. To measure these qualities, we evaluate captions based on both verification, in which a network is trained to predict if two captions are of the same person, and identification, in which a network is trained to compute a similarity score between an image and a caption.

To enforce consistency of evaluation, we use natural language processing techniques to evaluate captions rather than directly using the attributes used to generate our captions. This is to ensure results are derived from the captions being generated rather than just the attributes, which vary in number between methods and for some methods may not exist.

Verification

To perform verification, we fine-tune a DistilBERT model [60] to predict whether two captions describe the same person. We use the tokenizer and pretrained weights provided by Huggingface Transformers [61], which were trained using BookCorpus and

Caption Type	**Not Consistent**	**Consistent**
Gender only	73.39%	
CelebA attributes	81.15 ± .06%	81.70 ± .02%
CelebA + Shape attributes	81.16 ± .04%	82.60 ± .07%

Table 4.4: Verification accuracy for CelebA and shape label-based captions using consistent and inconsistent labels.

English Wikipedia. We use DistilBERT instead of the larger BERT model because our focus is on evaluating captions rather than producing a state-of-the-art text verification model, and DistilBERT can be more quickly and efficiently trained. To train the model, we randomly sample images from the training set with equal probability assigned to sampling two images of the same person and sampling two images of different people. We then predict both CelebA attributes and shape features using ResNet-18 models. ResNet-18 model as described in Section 3.1.1. To evaluate the model, we iterate through the CelebA validation set, matching each image with one of the same person or one of a different person with 50% probability. The model is trained for 3 epochs with a batch size of 64 using the Adam optimizer with weight decay.

We find that we can obtain reasonable results using CelebA captions, but these results are highly reliant on the gender attribute. With captions using only CelebA attributes, we achieve a verification accuracy of 81.15 ± .06% on the validation set (averaged over 3 runs). Using the identity-consistent verison of the dataset only improves validation accuracy to 81.70 ± .02%. Note that with a balanced gender distribution, a verification accuracy of 75 can be achieved simply by always predicting 0 when the genders don't match and always predict 1 otherwise – using captions only containing gender generated from predicted attribute labels, we are able to achieve a validation accuracy of 73.39%. With default CelebA attributes we are therefore only able to improve over this simple baseline by 7.76%. These results are shown in Table

4.4.

Captions including shape features improve upon this accuracy, though not by a huge amount. We find that shape labels generated on a per-image basis do not improve verification. However, while forcing CelebA captions to be consistent across identities only results in moderate improvement (+.55%), forcing the segment captions to be consistent improves accuracy by a further .90%. Note that, as previously mentioned, attribute captions do not substantially improve over a simple gender-based baseline so improvements of .90% are relatively large.

Identification

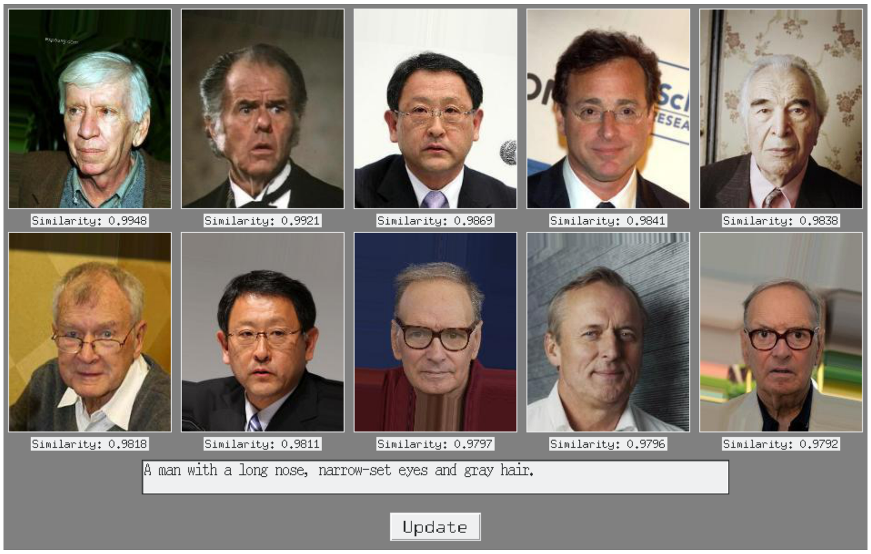
To perform identification, we use a contrastive technique somewhat similar to CLIP [62]. We map images and captions to a common representation space where we can use dot product to compute the similarity between an image and a caption, and for the identification task simply return the image or caption with the highest similarity score. We again use DistilBERT to generate base representations for the captions, and use ResNet-18 to generate base representations for the images. A two-layer fully connected network with a ReLU nonlinearity is then used to project these base representations to a smaller subspace in which the contrastive loss is performed. We use a size of 128 for the projected representations. For our contrastive loss, we the NT-Xent normalized cross-entropy loss [63] with a temperature parameter of 1. The loss is computed by comparing each element in a mini-batch to each other element in that same mini-batch, where the only similar representations should be an image and its associated caption. For evaluation, we use the CelebA validation set. Top-1 and top-10 accuracy are computed using the dot product between each image and each caption, then computing the number of images for which the correct caption is the most similar or among the 10 most similar captions.

	Not Consistent		**Consistent**	
Caption Type	Top-1	Top-10	Top-1	Top-10
CelebA attributes	43.44 ± 1.27%	94.73 ± .26%	40.74 ± 1.58%	94.04 ± .38%
CelebA + Shape	44.10 ± 1.23%	94.79 ± .44%	41.65 ± 1.07%	94.14 ± .46%

Table 4.5: Validation accuracy for CelebA and shape label-based captions using consistent and inconsistent labels.

We find that captions are reasonably capable of placing the correct image within the top few results, but not always consistent for returning the correct image first. With the default CelebA labels, we are able to achieve an average top-1 accuracy of 43.44% and an average top-10 accuracy of 94.73% with a batch size of 128. With consistent captions, we achieve a top-1 accuracy of 40.74% and a top-10 accuracy of 94.14%. We note that it is expected for the consistent captions to perform worse, because enforcing consistency intentionally throws away per-image biases which would make verification more difficult but identification easier. However, the accuracy difference is very small, indicating that these biases are not actually very useful for identification.

As shown in Table 4.5, captions using shape-based labels do not substantially improve upon the top-10 accuracy of a classifier using just CelebA attributes. There is a more significant difference in top-1 accuracy for both the consistent and non-consistent captions, but results are far more variable even when averaged over 5 runs. We are therefore unable to conclude if the shape-augmented captions are better than the plain CelebA captions for identifying a specific image. While shape-based labels do not seem to substantially improve upon identification results, they do provide additional options for text-based image lookup which are closer to a standard definition or “prominent features.” An example of this is shown in Figure 4.5. Note that for this application we modify use captions for which labels are discarded with 50% probability (if the gender label is discarded, pronouns are replaced with “they/them”). We



A man with a long nose, narrow-set eyes and gray hair.

Figure 4.5: Application for searching for images using text descriptions. The selected images are pulled from a pool of 2,000 validation images.

also use all segment features rather than just the ones with high information gain.

4.2.3 Summary

Based on the usefulness of captions for image description and lookup, we suggest two methods for evaluation: caption verification and caption identification. The former measures how well the captions describe what is unique about a person, whereas the latter measures how well the caption describes what is unique about a specific image. Our evaluation method allows comparison between different face captioning methods irrespective of how the captions where generated. We apply this to captions generated from CelebA labels as well as captions generated from a combination of CelebA labels and weak labels generated based on information calculated using semantic segmentation. Our weak labels are based on simple measurements computed from semantic segmentation data on frontalized images and do a reasonable job of approximating ground-truth labels. We find that our augmented captions result in an improvement in verification accuracy, though it is unclear if the captions improve identification accuracy. This aligns with our focus on recognizing prominent features which are important for verification.

Chapter 5

Conclusion and Future Work

Natural language face description is an interesting and important area of research, but existing methods do a poor job of capturing what humans recognize as prominent facial features. Prominent feature descriptions collected without restriction are widely varied (we collected 150 unique descriptions for a dataset of 205 people) and have little crossover with existing attribute datasets. The most commonly used attribute dataset, CelebA, is highly flawed and can be shown to be inconsistent or frequently wrong for at least 10 of the 40 attributes present in the dataset. The attributes that would be most useful for describing prominent features are not consistent across different images of the same person and in many cases do not accurately represent the feature they are describing. As a result of these issues, researchers should be cautious about making performance claims in regards to CelebA facial attribute classifiers. Future work should considering separating the most subjective attributes or even removing them from consideration entirely.

Furthermore, the existing methods for attribute prediction do a poor job of measuring their ability to perform imbalanced classification. The two metrics primarily used to compare CelebA performance, accuracy and balanced accuracy, can be optimized for imbalanced attributes without producing a classifier which is actually useful for predicting those attributes. The small scale of differences between attribute prediction methods further complicate comparison between models.

Because of these issues, we generate weak labels using semantic segmentation. While this approach has flaws – in particular, it struggles with non-frontal images – we are able to accurately capture a number of prominent facial features. In future work we will expand our weak labeling method to cover more features and investigate ways to better deal with issues with non-frontal images.

Although the existing data is flawed, we investigate the possibility of using attribute data to generate captions of face images. We find that these captions rely heavily on gender, but with consistent features and our weakly labeled prominent features can reasonably describe a person's unique features. With textual descriptions alone we are able to achieve a verification accuracy of 82.6%. Because there is little existing work in the area, it is currently unknown what "human-level" accuracy on this task is. We leave better interpretation of these accuracies to future work.

We find that captions can more reliably be used for identification, and take advantage of this to build a BERT-based system for text-based image lookup. Our lookup system can extract and utilize 54 attributes (40 from CelebA, 14 from prominent feature labels) for searching for faces.

There remain many open challenges to be dealt with in future work. Due to the subjectivity of prominent features, it is very difficult to collect accurate, unbiased labels. Binary labeling is problematic for many prominent features because it is unclear where the binary cutoff should be (e.g., it is not clear how "big" a nose should be before the image is labeled as "big nose"). There are also some features for which it is inherently problematic for the feature to be binary, and even more problematic to have labelers make "ground-truth" determinations about that feature.

Gender, for example, is not a binary feature despite being reduced to one by attribute recognition datasets and labelers with no knowledge of the subject are unqualified to make determinations about the subject's gender.

While weak labeling helps deal with labeling bias, weak labels are not without issues. Our weak label features define "prominent" relative to the dataset they are generated for, and as such inherit the biases of that dataset. CelebA disproportionately consists of white people, and as such a network trained using CelebA will learn to define "prominent features" largely as "features which are prominent for white people." This is a major obstacle to practical usage of our work, and is an important area for future work. Note that our caption evaluation method does not account for bias in the labeling, and as such bias must be investigated separately.

Using semantic segmentation is also highly flawed because segmentation can only capture 2D information. We found that in practice 3D models are not any more appropriate for prominent feature detection because existing methods for 3D face mesh prediction are designed primarily for alignment and as a result tend to "normalize" face meshes to make the most prominent features more average. Semantic segmentation similarly does not perfectly capture the size of each feature, but is in general able to do a better job because the problem is simpler and there is much more training data available. However, unlike segmentation, 3d-based features are not inherently limited and will likely be necessary to make further improvements in prominent feature detection.

Bibliography

[1] M. Sheehan and M. Nachman, "Morphological and population genomic evidence that human faces have evolved to signal individual identity," *Nature communications*, vol. 5, no. 4800, 2014.

[2] G. Rhodes, *Superportraits: Caricatures and recognition.* Psychology Press, 1997.

[3] 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.* IEEE, 2009, pp. 365–372.

[4] 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*, 2017, pp. 365–374.

[5] F. Taherkhani, N. M. Nasrabadi, and J. Dawson, "A deep face identification network enhanced by facial attributes prediction," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 553–560.

[6] M. M. Kalayeh and M. Shah, "On symbiosis of attribute prediction and semantic segmentation," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[7] 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)*. IEEE, 2017, pp. 17–24.

[8] 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*, 2018, pp. 8789–8797.

[9] 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*, vol. 28, no. 11, pp. 5464–5478, 2019.

[10] 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.

[11] D. Stap, M. Bleeker, S. Ibrahimi, and M. ter Hoeve, "Conditional image generation and manipulation for user-specified content," *arXiv preprint arXiv:2005.04909*, 2020.

[12] W. Xia, Y. Yang, J.-H. Xue, and B. Wu, "Tedigan: Text-guided diverse face image generation and manipulation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[13] C. Frowd, F. Skelton, C. Atherton, M. Pitchfork, V. Bruce, R. Atkins, C. Gannon, D. Ross, F. Young, L. Nelson, G. Hepton, A. McIntyre, and P. Hancock, "Understanding the multiframe caricature advantage for recognizing facial composites," *Visual Cognition*, vol. 20, no. 10, pp. 1215–1241, 2012.

[14] K. Lee, G. Byatt, and G. Rhodes, "Caricature effects, distinctiveness, and identification: Testing the face-space framework," *Psychological Science*, vol. 11, no. 5, pp. 379–385, 2000.

[15] N. Kumar, A. Berg, P. N. Belhumeur, and S. Nayar, "Describable visual attributes for face verification and image search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 10, pp. 1962–1977, 2011.

[16] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2015, pp. 34–42.

[17] R. Rothe, R. Timofte, and L. V. Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," *International Journal of Computer Vision*, vol. 126, no. 2-4, pp. 144–157, 2018.

[18] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[19] B. M. Smith, L. Zhang, J. Brandt, Z. Lin, and J. Yang, "Exemplar-based face parsing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3484–3491.

[20] A. D. Bagdanov, A. Del Bimbo, and I. Masi, “The florence 2d/3d hybrid face dataset,” in *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, 2011, pp. 79–80.

[21] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, “Face alignment across large poses: A 3d solution,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 146–155.

[22] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, “Panda: Pose aligned networks for deep attribute modeling,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1637–1644.

[23] N. Kumar, P. Belhumeur, and S. Nayar, “Facetracer: A search engine for large collections of images with faces,” in *European conference on computer vision*. Springer, 2008, pp. 340–353.

[24] 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*. Springer, 2016, pp. 19–35.

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

[26] 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*, vol. 31, 2017.

[27] 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*, 2017, pp. 6942–6950.

[28] 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*. International Joint Conferences on Artificial Intelligence Organization, 7 2018, pp. 733–740. [Online]. Available: https://doi.org/10.24963/ijcai.2018/102

[29] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Deep imbalanced learning for face recognition and attribute prediction," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 11, pp. 2781–2794, 2019.

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

[31] 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*, 2020, pp. 13 055–13 064.

[32] 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*, vol. 32, 2018.

[33] V. U. Prabhu, D. A. Yap, A. Wang, and J. Whaley, “Covering up bias in celeba-like datasets with markov blankets: A post-hoc cure for attribute prior avoidance,” *arXiv preprint arXiv:1907.12917*, 2019.

[34] Z. Wang, K. Qinami, I. C. Karakozis, K. Genova, P. Nair, K. Hata, and O. Russakovsky, “Towards fairness in visual recognition: Effective strategies for bias mitigation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8919–8928.

[35] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, “Men also like shopping: Reducing gender bias amplification using corpus-level constraints,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017.

[36] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 2048–2057.

[37] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, “Unified vision-language pre-training for image captioning and vqa,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 13 041–13 049.

[38] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, “The caltech-ucsd birds-200-2011 dataset,” 2011.

[39] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, “Microsoft coco captions: Data collection and evaluation server,” *arXiv preprint arXiv:1504.00325*, 2015.

[40] F. Yan, J. Kittler, and K. Mikolajczyk, "Person re-identification with vision and language," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 2136–2141.

[41] A. Gatt, M. Tanti, A. Muscat, P. Paggio, R. A. Farrugia, C. Borg, K. P. Camilleri, M. Rosner, and L. van der Plas, "Face2text: Collecting an annotated image description corpus for the generation of rich face descriptions," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[42] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "Face recognition from caption-based supervision," *International Journal of Computer Vision*, vol. 96, no. 1, pp. 64–82, 2012.

[43] Q. Huang, L. Yang, H. Huang, T. Wu, and D. Lin, "Caption-supervised face recognition: Training a state-of-the-art face model without manual annotation," in *European Conference on Computer Vision*. Springer, 2020, pp. 139–155.

[44] Y. Feng, L. Ma, W. Liu, and J. Luo, "Unsupervised image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4125–4134.

[45] T. Wang, T. Zhang, and B. Lovell, "Faces a la carte: Text-to-face generation via attribute disentanglement," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3380–3388.

[46] 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*, 2016, pp. 770–778.

[47] 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)*. IEEE, 2016, pp. 3239–3243.

[48] J. Li, F. Zhao, J. Feng, S. Roy, S. Yan, and T. Sim, “Landmark free face attribute prediction,” *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4651–4662, 2018.

[49] 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.

[50] 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*, 2017, pp. 1851–1860.

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

[52] J. M. Johnson and T. M. Khoshgoftaar, “Survey on deep learning with class imbalance,” *Journal of Big Data*, vol. 6, no. 1, p. 27, 2019.

[53] R. Rothe, R. Timofte, and L. Van Gool, “Deep expectation of real and apparent age from a single image without facial landmarks,” *International Journal of Computer Vision*, vol. 126, no. 2, pp. 144–157, 2018.

[54] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[55] N. Ruiz, E. Chong, and J. M. Rehg, “Fine-grained head pose estimation without keypoints,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[56] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition.* Ieee, 2009, pp. 248–255.

[57] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention.* Springer, 2015, pp. 234–241.

[58] J. R. Quinlan, “Induction of decision trees,” *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[59] H. Zhou, J. Liu, Z. Liu, Y. Liu, and X. Wang, “Rotate-and-render: Unsupervised photorealistic face rotation from single-view images,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[60] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter,” *arXiv preprint arXiv:1910.01108*, 2019.

[61] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, “Transformers: State-of-the-art natural language processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.* Online: Association

for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[62] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” *arXiv preprint arXiv:2103.00020*, 2021.

[63] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International conference on machine learning.* PMLR, 2020, pp. 1597–1607.