Improving Facial Attribute Recognition with Attribute Localization Data

Nathan Thom and Emily M. Hand
University of Nevada, Reno
1664 N. Virginia St.
Reno, Nevada 89557
nathanthom@nevada.unr.edu, emhand@unr.edu

Abstract

Facial attribute recognition has gained popularity in recent years with the advent of deep learning. Attributes are describable features, with facial attributes including features such as gender, hair color, facial hair, etc. Despite its popularity, there are only two large-scale datasets available for this problem – CelebA and LFWA – consisting mostly of celebrities posed on the red carpet. This leads to unequal representation of each labeled attribute (e.g. many young people, few bald people, etc.). Because of this, many current methods rely on correlations between attributes in the training set rather than identifying each attribute independently [6]. Learning these misleading correlations leads to poor generalization and biased prediction. In this work, we introduce a novel multi-task learning architecture for the task of facial attribute recognition. Prediction is combined with localization in an effort to focus the model on the presence of each attribute, rather than correlations between attributes. Localization is achieved with a novel, weak labeling approach for semantic segmentation of facial attributes.

1. Introduction

Facial attribute recognition was first introduced as a means to improve face verification performance [11, 12, 13]. Face verification is the problem of identifying whether or not two images contain the same person. Now a field in its own right, facial attribute recognition focuses on classifying human describable features of faces in images or video. Prior to deep learning, research in facial attribute recognition included the FaceTracer face search engine [11], simile classifiers on the PubFig dataset [12], and multi-label classification on the Labeled Faces in the Wild (LFW) dataset [7].

In 2015, deep learning methods became popular for facial attribute recognition, with the introduction of two large-scale datasets for the problem: CelebA and LFWA [16]. With the introduction of these large scale datasets, the number of deep learning methods for the problem of attribute recognition exploded. Although CelebA allowed for significant progress to be made in the field, it has been shown to have significant label imbalance, with many of the methods based on this dataset relying on correlations between unbalanced features [21, 5]. To address this issue, we propose a joint learning architecture in which attribute recognition is combined with semantic segmentation – a task that is independent of inter-class correlations.

Semantic segmentation is the problem of classifying every pixel in an image as belonging to one or multiple classes. State of the art methods for semantic segmentation utilize convolutional neural networks (CNNs) and seek to identify a single class for each pixel in an image [17, 19, 4]. Very few works address the problem of semantic segmentation of faces. Kalayeh et al. propose segmenting the face into parts for improved attribute recognition [9]. This however still differs from our approach, which segments faces according to where each attribute occurs on the face. Thus, we have generated a novel, weak labeling of attribute segments for the CelebA dataset. This enables learning of attribute localization alongside attribute recognition.

The proposed work introduces a novel technique for facial attribute recognition. A basic facial attribute recognition model is strengthened with additional supervision from a weakly labeled semantic segmentation task. Segmentation labels are generated essentially for free by automatically extracting facial landmarks, then a rule-based system uses landmarks to label the portions of input images where attributes occur. Prior work has shown that many state-of-the-art algorithms rely very heavily on attribute correlations, rather than the actual presence of an attribute [21, 5]. We address this problem through the combination of weakly supervised semantic segmentation and attribute recognition in one learning framework. By generating weak segmentation labels for each attribute, our method learns where to look for an attribute as well as what to look for when recognizing attributes of a face. We show that this multi-task

framework leads to an improved representation of facial attributes which does not merely rely on correlations between classes.

To summarize, this work's contributions include:

- AttParseNet: a multi-task CNN for simultaneous attribute localization and recognition using a weakly labeled training approach.
- A framework for generating semantic segmentation labels in the context of facial attributes
- Weak attribute segments for the full CelebA dataset, to be released with the publication of this work.

2. Related Work

The proposed research combines work in semantic segmentation and attribute recognition. We detail the relevant literature in the following sections.

2.1. Semantic Segmentation

In semantic segmentation, the goal is to assign a class label to every pixel in an image, effectively segmenting it into its parts. Face parsing is a form of semantic segmentation that separates the face into its parts (eyebrows, mouth, nose, etc.). Traditionally, Conditional Random Fields were used by all state-of-the-art methods for face parsing [26, 8, 24]. As in many other fields, deep learning became the new state-of-the-art in face parsing and semantic segmentation.

With the introduction of Fully Convolutional Networks in [23], deep learning became the go-to method for semantic segmentation. There is less work on face parsing than semantic segmentation, with the focus in most face processing work on landmark localization rather than segmenting faces into their parts. [15] combines CNNs with Conditional Random Fields for improved face parsing. [18] uses a hierarchical deep learning approach focusing on parsing faces with partial occlusions. More recently, [22] combines facial alignment with segmentation, using a shared representation to improve learning of both tasks for the purpose of virtual makeup and face swapping. Attention has also been introduced as a way to improve semantic segmentation when working with multi-scale images [2].

The proposed methods differ from prior work in several ways. First, we seek to parse faces according to attributes, rather than into parts. While traditional semantic segmentation techniques focus on a multi-class labeling problem, our work seeks to address a multi-label classification problem, assigning a set of attributes to each pixel in an image. Second, our work takes advantage of weak attribute segments by utilizing a rule-based method which generates segments from facial landmarks and attribute labels. This constrains the attribute model to focus on regions of interest, keeping it from taking advantage of attribute relationships that may only be present in the training data. Additionally, unlike most semantic segmentation techniques, our weakly labeled facial attribute segmentation technique requires no human labeling.

Figure 1. Sample images from the CelebA dataset [16].

2.2. Attribute Recognition

Facial attribute recognition was first introduced by Kumar et al. in [11]. The same group later showed that attributes were useful for search and retrieval as well as face verification [12, 13]. in 2015, Liu et al. introduced the large-scale benchmark dataset CelebA, containing over 200,000 images each labeled with 40 binary attributes [16].The introduction of CelebA was a significant milestone for the field because its size enabled the use of deep learning methods, which was previously not possible due to lack of data. The dataset is still widely used today. Sample images from CelebA are shown in Figure 1.

Along with CelebA, Liu et al. introduced a method for face localization and attribute recognition that involves two networks: LNet and ANet [16]. LNet performs localization for faces with weak attribute supervision, and ANet uses the localized face to predict facial attributes. Both LNet and ANet are built on the AlexNet [10] architecture pre-trained on the ImageNet dataset [3]. LNet is then fine-tuned to predict facial attributes from the full body images in CelebA. This training enables LNet to localize the face in a given image. Once the face is localized by LNet, ANet is trained to recognize attributes from the cropped face image. This two-network scheme produced impressive results on CelebA. The proposed method significantly differs from [16]. We introduce a rule-based method for generating weak attribute segment labels for each image, without the need for any human labeling. he proposed model, AttParseNet, is trained directly from CelebA, starting with randomly initialized weights, whereas most previous methods are pre-trained on ImageNet or other large-scale datasets. Additionally, AttParseNet jointly localizes and recongizes attributes in

one model using a weak semantic segmentation task during training and no segments at test time.

[21, 5, 25] recently introduced methods to combat label imbalance for the problem of attribute recognition. The Mixed Objective Optimization Network (MOON) from [21] addressed label imbalance by calculating source and target distributions for each attribute and applying a weight to the backpropagation to adjust for the difference between the distributions. As a follow-up to MOON, [5] introduced a method called "Selective Learning" where balancing is performed at the batch level by weighing underrepresented classes and sampling from over-represented ones. [25] introduces an automated technique to remove label noise from CelebA, building representative sets with which to compare each sample, achieving very high recall on noisy sample removal. The proposed work attempts to reduce the effects of label imbalance in the CelebA dataset by focusing on the problem of attribute localization using weakly labeled semantic segmentation as an additional task. Combining semantic segmentation with attribute recognition forces the model to focus on regions of interest when predicting attributes, and provides extra supervision essentially for free (without requiring human labels).

There is only one work that is closely related to the proposed methods. It comines semantic segmentation with attribute recognition, but focuses on segmenting parts using hand-labeled data for training [9, 14]. The Helen dataset is labeled with face parts at the pixel level, and is used in [9] to focus the network on areas of interest for each attribute using pooling and gating mechanisms. The proposed work differs significantly from that of [9] in that we use a multi-task learning framework to perform attribute segmentation and recognition from weakly labeled data. We generate weak attribute segments for all of CelebA with no hand labeling, and combine the tasks of attribute segmentation and recognition in one model forcing the network to both localize and recognize facial attributes. We highlight that our method does not perform standard face parsing. Rather than parsing the face into parts, our method parses the face into attributes, producing a map of attribute locations, rather than part locations.

3. Proposed Methods

The proposed method consists of two main parts: the generation of weak segmentation labels, and the multi-task learning framework, which combines attribute segmentation and recognition. We detail both in the following subsections.

3.1. Weak Segment Generation

Teaching the model to localize facial attributes is facilitated by semantic segmentation labels. This form of labeling assigns classes to each pixel in an input image. For the scope of this work, the pixels of each image in CelebA are labeled with the presence or absence of 40 attribute classes. The binary labels for these classes are provided along with the CelebA dataset. Example attribute classes are *smiling*, *wavy hair*, *young*, etc. Our semantic segmentation labels are represented as masks of the same height and width as the input images and a depth of 40 channels (one for each attribute class). Segment masks have a value of 1 in areas where the attribute is present and 0 everywhere else. Hand-labeling this data is expensive and slow, so we opt to automate the process by introducing a weak labeling strategy, which requires no human supervision.

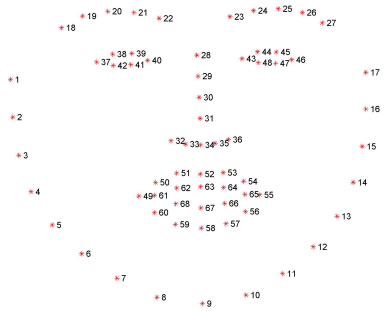

Figure 2. Layout of facial landmarks extracted from OpenCV and OpenFace.

Generation of segment labels begins by extracting a set of facial landmarks from each image in CelebA. Figure 2 shows the layout of the 68 facial landmarks that are used. We utilize the OpenCV and OpenFace landmark detectors to extract these points [1, 27]. This technique yields fiducial points for over 99% of CelebA. The remaining images are hand-labeled with landmarks.

The set of collected facial landmarks are used to define the regions where attributes most commonly occur on a face. For example, we assume that the attribute *smiling* occurs in the same region of the face as the mouth. This region is enclosed by landmark points $\{49\text{-}60\}$. Similar regions are defined for all 40 attributes. Combining this information with the attribute labels in CelebA enables a nearly automatic system for producing segmentation masks for the entire dataset. This method is significant because it provides a framework for producing additional layers of supervision on arbitrary attribute recognition tasks. In addition, the data is generated with very little overhead.

We generate weak segmentation masks by forming and filling polygons (setting pixel intensities to 1) on black images (all pixels set to 0). The polygons are constructed from the extracted landmarks and represent the area of the input image where an attribute is present. 10 base regions are used in combination to create these shapes. The base regions are *below chin*, *chin*, *cheeks*, *mouth*, *above mouth*, *nose*, *eyes*, *eyebrows*, *ears*, and *top of head*. The *chin*, *mouth*, *nose*, *eyes* and *eyebrows* regions are precise because they are de-

Figure 3. Examples of the 10 base regions used to generate weak semantic segmentation labels. These regions are overlayed with the original images for visualization purposes. The segment regions are show in blue and landmark points are red.

fined directly from the 68 landmark points. The remaining 5 regions are established by combining these precise regions with information about facial geometry. For example, the *top of head* region is created by using landmarks from the eyebrows and information about facial geometry, since no landmarks for the forehead are given. Figure 3 shows the different regions used in the generation of attribute segments.

We consider the segmentation labels to be weak for two reasons: 1) our rule-based method for generating segments is built on automated facial landmark extraction, so the landmarks and regions may be imprecise, and 2) there are several attributes for which the physical manifestation is unclear, and so the proposed segments may not provide adequate coverage. We provide several examples to clarify the two types of weak labels in our segmentation work. For type-1 segments, if mouth landmarks are misaligned then the mouth segments will be incorrect. Additionally, there are no landmarks for hair, so all hair related attributes (e.g. *brown hair*, *wavy hair*) have rough segments from the *top of head* region. For type-2 segments, there is much debate in the field of expression and micro-expression recognition as to what indicates a smile: the mouth or other deformations of the face around the eyes. In this work, we simply assume the mouth is responsible for mouth-related attributes and so we may be missing out on other cues in the faces.

3.2. Attribute Segmentation and Recognition

Once the weakly labeled attribute segments have been generated, the next step is to build a model which learns to recognize attributes. Attribute recognition and segmentation are learned jointly in a model that we call AttParseNet. The task of semantic segmentation is used to improve our model's attribute recognition accuracy and generalizability.

The proposed multitask attribute segmentation and recognition model is an eight-layer CNN. The architecture for the CNN is shown in Figure 4. The model consists of six convolution layers, the first using filters of size 7x7, and the remaining layers using filters of size 3x3. The number of filters in each layer is as follows: 75, 200, 300, 512, 512 and 40. Max pooling is performed after the first convolution layer. After the final convolution layer, the model produces 40 features maps each of size 96x76. This can be interpreted as 40 segmentation outputs, one for each attribute in CelebA. The generated segment masks are the same size as the input images. In order to compare the masks with the feature maps from the final convolution layer, we must down-sample the weak segment masks to the size 96x76. We use nearest neighbor interpolation to retain the binary nature of the pixels. AttParseNet jointly learns segmentation and recognition. The segmentation loss is formulated as mean squared error (MSE) between the output feature maps and segment masks.

For the recognition task, the feature maps are flattened and passed into one fully connected layer, resulting in a final 40-dimensional output (one for each attribute). This layer is used to perform attribute recognition, based on the segment feature maps that result from the six convolution layers. The additional supervision from the semantic segmentation task causes AttParseNet to activate in relevant regions more consistently. The 40-dimensional output is compared with the attribute labels for the image using a binary cross-entropy (BCE) loss with logits.

In order for AttParseNet to learn from both the segmentation and recognition tasks in one framework, both tasks have their own loss functions. Joint learning is achieved by computing the summed total of both loss functions and back propagating according to both errors. The weak semantic segmentation task in AttParseNet provides an added level of supervision to the problem of attribute recognition for free. By "free" we mean that there is a very small amount of human labeling required, and the segments are generated using facial landmark points and weakly labeled using the image-level attribute labels provided with CelebA. The segmentation task provides an additional 291,840 labels for each image (40 feature maps, each of size 96x76). Adding weakly labeled semantic segmentation to AttParseNet forces the model to activate on regions of interest when learning attribute representations, which leads to a more robust and generalizable attribute model. We showcase this in our experiments. **It is important to note that the weak segments are used only at training time, and are not needed during testing.** The segmentation task is used to localize attributes during training, allowing the model to focus on regions of interest at test time without landmarking or segmentation.

4. Experiments and Results

We begin experimentation on the CelebA dataset. CelebA consists of 202,600 images, each image is labeled

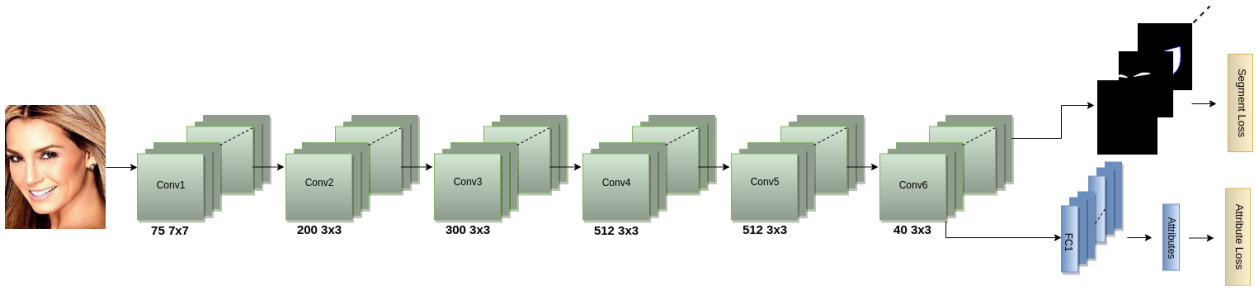

Figure 4. Our multi-task learning architecture. Input of an image is provided and is passed through 6 shared convolutional layers. The network outputs segmentation masks and attribute predictions.

with 40 attributes. It should also be noted that the dataset features a cropped and aligned set of images and full body, unaligned images. We utilize the full body, unaligned images for training of AttParseNet. We use our extracted landmark points to crop the images about the face. Finally, our model takes images with dimensions 218x178 as input, so the cropped images and segmentation labels are resized to dimensions of 218x176 and 96x76 respectively.

AttParseNet is trained on the unaligned, cropped images from the CelebA training split – without training of any kind on an external set of data. Our training images differ significantly from the CelebA aligned images because there is more pose variance and the images are more focused on the face region. The proposed work is trained in two stages. First, the model is trained for 10 epochs and network weight updates are based only on the MSE loss from the weak segmentation task. Next, AttParseNet is trained for 22 epochs on both the segmentation and recognition tasks. During the second phase of training MSE and BCE loss functions are summed. Network weights are updated with the Adam optimizer and a learning rate of .001. AttParseNet uses the weakly labeled segments for training only. **During validation and test, segment labels are not used**. The idea is that the segmentation learning task allows AttParseNet to focus on localization as well as recognition of attributes so that at test time, localization is a part of the model and segments are not needed. Therefore, there is no need for landmark extraction at test time. In order to encourage future research in this direction, we have made the weak attribute segments for CelebA publicly available.

For comparison, we train another model which has an identical architecture to AttParseNet, as well as the same hyperparameters. The only differences are: the segmentation learning task is not used and training is completed with the aligned dataset for fairness. Training on the aligned set of images is necessary because weak segment supervision from image-level attribute labels provides implicit alignment. We use this network for comparison so as to understand how learning localization alongside attribute recognition affects AttParseNet's ability to recognize a broad-range of facial attributes, in addition to its generalizability to various unseen data.

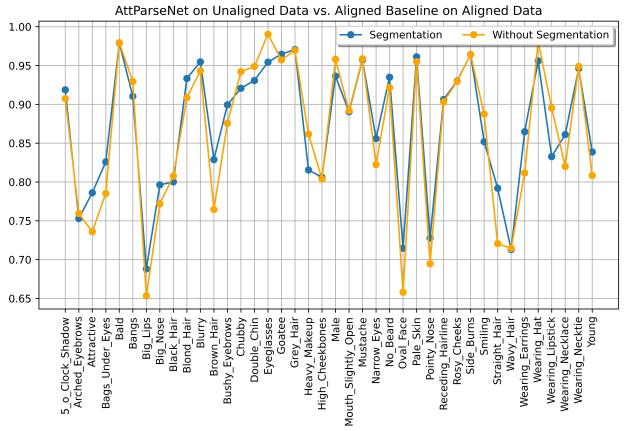

Figure 5. Average accuracy achieved on each facial attribute for the proposed architecture and a baseline model. The models are evaluated on the unaligned and aligned data sets respectively. AttParseNet is trained with the weak semantic segmentation task.

Both networks are implemented in PyTorch [20]. The data is split according to the training, validation and test provided with CelebA. We train our model on a two GTX-1080 TI GPUs. Overfitting is avoided by training only until loss on the training set is comparable to loss on the validation set.

The aligned baseline (without segmentation) model achieves an average attribute accuracy of 86% on the aligned test set of CelebA, whereas AttParseNet achieves an average of 87% on the unaligned test set. Although small, this improvement is substantial when one considers that the accuracy is averaged over 40 attributes. Figure 5 shows the average accuracy achieved by both networks for each attribute on the test split. Our results show that learning to localize features is beneficial for attribute recognition in general. We see improved accuracy in over 50% of attribute classes. Half of the attributes which do not see improved accuracy occur in very rough segments, as landmarks outside of the face are not precisely extracted. To clarify, landmarks on the outside of the face are estimated through the combination of exact landmarks (see Figure 2) and facial

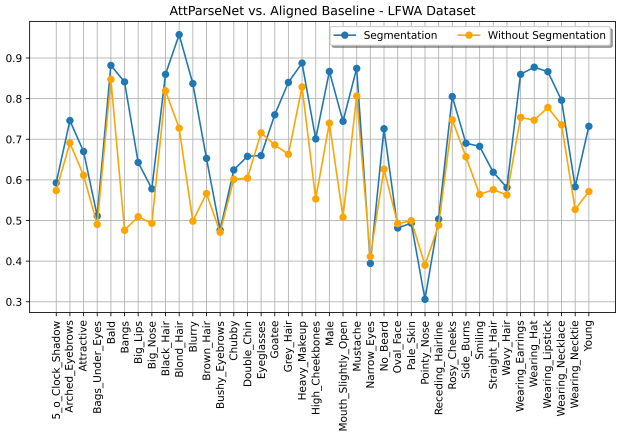

Figure 6. Average accuracy achieved on each facial attribute for the proposed architecture and a baseline model. AttParseNet is trained with the weak semantic segmentation task.

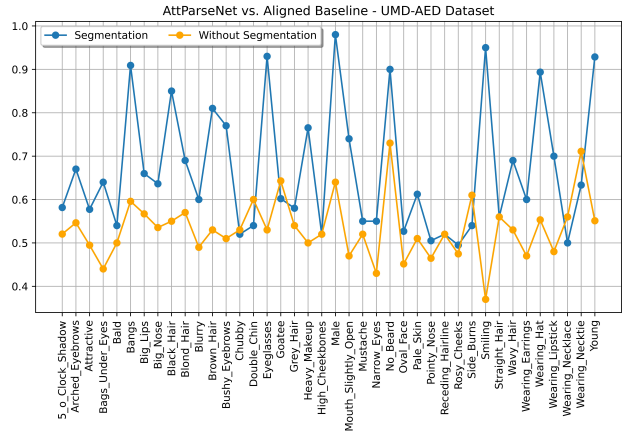

Figure 7. Average accuracy achieved on each facial attribute for the proposed architecture and a baseline model. AttParseNet is trained with the weak semantic segmentation task.

geometry. We show that this loss of accuracy is likely due to our cropping method, which often removes data on the outside edges of a face.

To exemplify the improved generalizability of AttParseNet, we test the dataset of the LFWA and UMD-AED datasets. LFWA is a very significant dataset in the field of facial attribute recognition and UMD-AED boasts almost perfect balance of labels between attribute classes, making both relevant sources for testing data. Tests are completed by collecting predictions from AttParseNet and the baseline model for all data in each dataset, then accuracy is calculated based provided labels.

LFWA is examined first. See Figure 6. We note that all attribute classes are see increased performance besides *eyeglasses*, *narrow eyes*, *oval face* and *pale skin*. The difference of accuracy on each of these attributes is minor as it is less than 1% different, while some attributes are recognized by AttParseNet as much as 30% more effectively. Next, results on the UMD-AED dataset are analyzed. Accuracy for this trial is shown in Figure 7. Here we see improvement on all attributes besides *chubby*, *double chin*, *goatee*, *Sideburns*, *wearing necklace* and *wearing necktie*. Once again, each of the attributes that are not improved upon show less than 1% difference of accuracy, on average. This being said, performance differences for some classes are shown to be separated by nearly 40%. In particular we see significant improvements in several attributes that require precise localization, including *bangs*, *eyeglasses*, *black hair*, *smiling*, *no beard* and *wearing hat*.

In this work we make the assumption that some attributes such as *heavy makeup*, *chubby*, and *pale skin* come from all regions in the face. However, where exactly these attributes manifest themselves in a face image is unclear. Further research is needed to determine where exactly certain attributes come from. We leave this and the improvement of weak segmentation labels for future work.

5. Conclusions

In this paper we introduce a new method for facial attribute recognition from images, which we call AttParseNet. Our proposed method adds weakly labeled semantic segmentation of attributes as an additional level of supervision in the attribute recognition network. We also introduce a rule-based method for generating weakly labeled facial attribute segments based on landmark points. Using these weakly labeled attribute segments we are able to add a segmentation loss to the facial attribute recognition model, in addition to the attribute recognition loss. Combining these two learning tasks in a single network results in improved facial attribute recognition and generalizability of our model on unseen data. We demonstrate the effectiveness of our method, comparing AttParseNet with the a baseline model that has the same network architecture, but is trained without the segmentation task. AttParseNet shows significant improvements over the baseline on many attributes that require precise localization. These improvements suggest that previous state-of-the-art methods were taking advantage of attribute correlations in the datasets, and therefore were not learning a true representation for the attributes. AttParseNet is able to take advantage of weakly labeled segmentation data to better localize and recognize facial attributes, requiring no facial landmarking at test time. **We emphasize that the proposed work required very little hand-labeling and no new data was collected.** Rather, we introduced a rule-based method to create weak semantic segmentation labels for added supervision in the task of attribute recognition. Even with very weak segment labels, AttParseNet is able to improve over the baseline method of attribute recognition without segmentation.

Future work consists of further refining the weak segmentation labels as well as a more detailed study of how attributes manifest themselves in the face.

References

[1] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[2] L. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. *CoRR*, abs/1511.03339, 2015.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[5] E. M. Hand, C. Castillo, and R. Chellappa. Doing the best we can with what we have: Multi-label balancing with selective learning for attribute prediction. AAAI, 2018.

[6] E. M. Hand and R. Chellappa. Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification. In *AAAI*, pages 4068–4074, 2017.

[7] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[8] A. Kae, K. Sohn, H. Lee, and E. Learned-Miller. Augmenting crfs with boltzmann machine shape priors for image labeling. 06 2013.

[9] M. M. Kalayeh, B. Gong, and M. Shah. Improving facial attribute prediction using semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[11] N. Kumar, P. Belhumeur, and S. Nayar. Facetracer: A search engine for large collections of images with faces. In *European Conference on Computer Vision*, pages 340–353. Springer, 2008.

[12] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Attribute and simile classifiers for face verification. In *International Conference on Computer Vision*, pages 365–372. IEEE, 2009.

[13] 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*, 33(10):1962–1977, 2011.

[14] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *European conference on computer vision*, pages 679–692. Springer, 2012.

[15] S. Liu, J. Yang, C. Huang, and M.-H. Yang. Multi-objective convolutional learning for face labeling. pages 3451–3459, 06 2015.

[16] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

[17] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[18] P. Luo, X. Wang, and X. Tang. Hierarchical face parsing via deep learning. pages 2480–2487, 06 2012.

[19] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. *CoRR*, abs/1505.04366, 2015.

[20] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

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

[22] Z. Shao, S. Ding, Y. Zhao, Q. Zhang, and L. Ma. Learning deep representation from coarse to fine for face alignment. 07 2016.

[23] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, April 2017.

[24] B. Smith, l. Zhang, J. Brandt, Z. Lin, and J. Yang. Exemplar-based face parsing. pages 3484–3491, 06 2013.

[25] J. Speth and E. M. Hand. Automated label noise identification for facial attribute recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 25–28, 2019.

[26] J. Warrell and S. Prince. Labelfaces: Parsing facial features by multiclass labeling with an epitome prior. pages 2481–2484, 11 2009.

[27] A. Zadeh, T. Baltrusaitis, and L. Morency. Deep constrained local models for facial landmark detection. *CoRR*, abs/1611.08657, 2016.