Within-Network Ensemble for Face Attributes Classification

Sara Atito Ali Ahmed[0000−0002−7576−5791] and
Berrin Yanikoglu[0000−0001−7403−7592]

Sabanci University, Istanbul, Turkey 34956
https://www.overleaf.com/2469964584kwwnghsrdwqw
{saraatito,berrin}@sabanciuniv.edu

Abstract. Face attributes classification is drawing attention as a research topic with applications in multiple domains, such as video surveillance and social media analysis. In this work, we propose to train attributes in groups based on their localization (head, eyes, nose, cheek, mouth, shoulder, and general areas) in an end-to-end framework considering the correlations between the different attributes. Furthermore, a novel ensemble learning technique is introduced within the network itself that reduces the time of training compared to ensemble of several models. Our approach outperforms the state-of-the-art of the attributes with an average improvement of almost 0.60% and 0.48% points, on the public CELEBA and LFWA datasets, respectively.

Keywords: Face Attributes Classification · Deep Learning · Multi-task Learning · Multi-label Classification · Ensemble Learning.

1 Introduction

Attribute classifiers have been drawing attention in zero-shot or few-shot learning problems where classes share attributes among them and can thus be recognized with zero or a few samples. Face attribute in particular has been a focus [13, 5, 17, 6, 7], as describing facial attributes has useful applications such as attribute-based search. Previously, work on face attribute classification approaches were based on handcrafted representations, as in [11, 3, 12]. This kind of approaches are prone to failing when presented different variations of face images and in unconstrained backgrounds. Recently, researchers tackle this task using deep learning, which has resulted in huge performance leaps in several domains [18, 19, 22, 13, 16, 21]. Liu et al. [13] use two cascaded convolutional neural networks (CNNs), for face localization (LNet) and attributes prediction (ANet). Each attribute classifier is trained independently where the last fully connected layer is replaced with a support vector machine classifier. Similarly in Zhong et al. [21], attribute prediction is accomplished by leveraging different levels of CNNs.

Lately, the task is shifted to be a multi-task learning (MTL) problem by training attributes in groups, mainly to speed up the training process and reduce overfitting. Yet, only few works address the relationship between different facial

attributes [7, 1, 6]. Hand and Chellapa's work divides the attributes into nine groups and train a CNN consisting of three convolutional sub-networks and two multi-layer perceptrons [7]. The first two convolutional sub-networks are shared for all of the classifiers and the rest of the network is independent for each group. They also compare their results to the results of classifiers trained independently for each attribute and show the advantage of grouping attributes together. Atito and Yanikoglu use the multi-task learning paradigm, where attributes that are grouped based on their location, share separate layers [1]. Learning is done in two-stages: first by directing the attention of each network to the area of interest and then fine-tuning the networks. In Han et al. [6], attributes are grouped into ordinal vs. nominal attributes, where nominal attributes usually have two or more classes and there is no intrinsic ordering among the categories, like race and gender. The attributes are jointly estimated by training a convolutional neural network that consists of some shared layers among all the attributes and category-specific layers for heterogeneous attributes.

In this work, we propose an end-to-end network where all of the attributes are trained at once in a multi-label learning scenario. An extra layer along with a combined objective function are added to the network to capture the relation between the attributes. Furthermore, a novel ensemble technique is introduced.

The main contributions are summarized as follows. (1) We use an end-to-end deep learning framework for face attribute classification, capturing the correlation among attributes with an extra layer that is trained at the same time with the first one. (2) We propose a novel within-network ensemble technique. (3) We obtain state-of-the-art results on both the CELEBA and LFWA datasets.

2 Proposed Approach

In this paper, we approached the face attributes classification problem in a multi-label/multi-task fashion using an end-to-end framework. In Sec. 2.1, we trained our base system in a multi-label fashion by sharing the network layers among all of the attributes. While in Sec. 2.2, we introduced groups and attributes specific layers for distinct feature extraction. In Sec. 2.3, an extra layer is embedded to the architecture to capture the relation between different attributes. Finally, in Sec. 2.4, a novel ensemble approach within the architecture itself is introduced.

Training a large deep learning network from scratch is time consuming and needs tremendous amount of training data. Therefore, all of our proposed architectures are based on fine-tuning a pre-trained model, namely the ResNet-50 network [8] which is the first place winner of the (ILSVRC) 2015 classification competition with top-5 error rate of 3.57%, trained on a dataset with 1.2 million hand-labeled images of 1,000 different object classes.

2.1 Base System

Multi-Task learning has already shown a significant success in different applications like face detection, facial landmarks annotation, pose estimation, and traffic flow prediction [15, 20, 10, 14].

In this work, we use MTL such that all the attributes are trained at once, using the same shared layers. To match the output of ResNet-50 network with our task, the output layer is replaced with 40 output units (one for each attribute) and use the cross-entropy loss function to measure the discrepancy between the expected and actual attribute values.

The multi-task approach not only saves on the training time, but the shared network is also more robust to overfitting, according to our experimental results. Intuitively, the model is forced to learn a general representation that captures all of the specified tasks which less the chance of overfitting. Similar findings are also reported in [2] and attributed to the regularization effect obtained by sharing weights for multiple tasks.

Group	**Attributes**
Head	(1) Black Hair (2) Blond Hair (3) Brown Hair (4) Gray Hair (5) Bald (6) Bangs (7) Straight Hair (8) Wavy Hair (9) Receding Hairline (10) Hat
Eyes	(11) Arched Eyebrows (12) Narrow Eyes (13) Bushy Eyebrows (14) Bags Under Eyes (15) Eyeglasses
Nose	(16) Big Nose (17) Pointy Nose
Mouth	(18) Big Lips (19) Smiling (20) Mustache (21) Wearing Lipstick (22) Mouth Slightly Open
Cheek	(23) 5 O-clock Shadow (24) Rosy Cheeks (25) Goatee (26) High Cheekbones (27) No Beard (28) Sideburns
Shoulder	(29) Double Chin (30) Wearing Necklace (31) Wearing Necktie
General	(32) Attractive (33) Blurry (34) Chubby (35) Young (36) Male (37) Pale Skin (38) Oval Face (39) Heavy Makeup, (40) Earrings

Table 1. Grouping attributes based on their relative location.

2.2 Multi-Task Learning with Attribute Grouping

When all the layers are shared in a simple multi-task learning approach, the resulting network may be overly constrained. Therefore, we added a residual block for each group of attributes, after the last residual network block (res5b), as well as few layers for each attribute. This architecture is shown in the dashed part of Fig. 1.

For grouping, the 40 attributes defined for the CELEBA and LFWA datasets are divided into 7 groups based on their localization (head, eyes, nose, cheeks, mouth, shoulder, and general areas) as shown in Table 1.

In the rest of the paper, we discuss our improvement to the multi-task learning network described thus far.

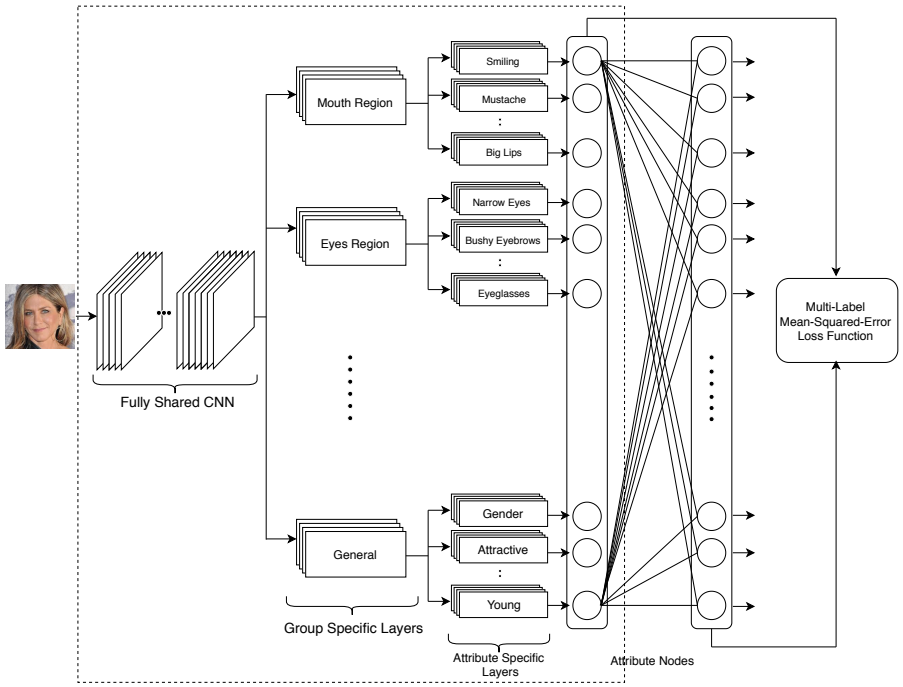

Fig. 1. End-to-end architecture for face attributes classification.

2.3 End-to-End Network

Neither the basic, nor the multi-task architectures so far take into account the correlations among attributes.

In previous work, correlations among facial attributes are learned and exploited by using a separate network or learning phase. In this work we add another fully connected layer with 40 output nodes to the network described in Section 2.2, for simplicity and end-to-end training. The resulting architecture is shown in Fig. 1, where the last layer aims to pick the most suitable predictions based on the predictions in the previous layer, by learning the correlations between the attributes.

The multi-label mean-squared-error loss used in this network consists of two terms, one for each of the last two layers. Specifically, for a given input image and A attributes, the loss function is denoted as shown in Equation 1, where $\hat{y}_1[a]$ and $\hat{y}_2[a]$ denote the output for attribute a, in the last two layers:

$$loss = \sum_{a=1}^{A} \left(y[a] - \hat{y}_1[a]\right)^2 + \left(y[a] - \hat{y}_2[a]\right)^2 \tag{1}$$

In this architecture, mean-squared-error loss is used instead of cross-entropy loss, with target values of $\{-1, 1\}$, since we aim to capture attribute correlations with the last layer weights.

2.4 Within-Network Ensemble

Ensemble approaches are very important in reducing over-fitting and they are used more and more to improving the performance of deep learning systems. However, forming ensembles from deep learning systems is very costly, as training often takes long hours or days.

To reduce the time to build the base classifiers forming the ensemble and inspired by the improved results with the end-to-end architecture with two output layers, we trained an ensemble all at once, within a single network.

The architecture illustrated in Fig. 2 shows the main idea behind our approach. Assuming that we have a classification/regression task with N outputs (here the 40 binary attribute nodes), we branch a fully connected layer with N output nodes after every several layers and include their error in the global loss function. During testing, the outputs of these branches are treated as separate base classifier outputs and averaged to obtain the final output.

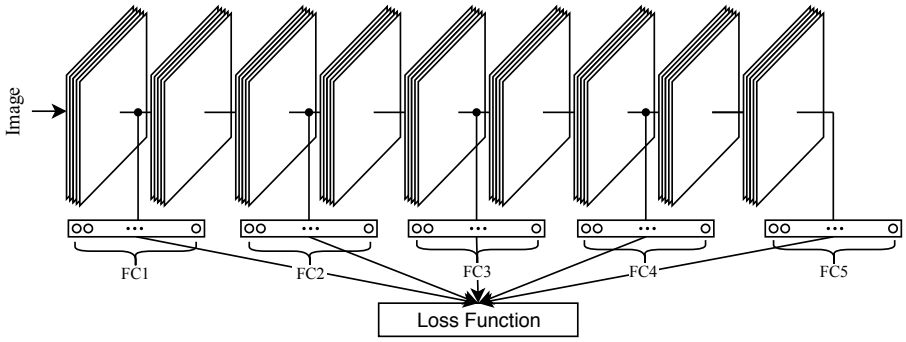

Fig. 2. A basic architecture of within-network ensemble approach, with 5 output layers.

In this work, we have constructed the ensemble with 5 such branches, each with 40 output nodes. The training of the network for one epoch on the LFWA dataset took approximately 18 minutes, compared to 16 minutes with the end-to-end network.

Notice that the base classifiers formed in this fashion use progressively more complex features and the training is much faster compared to training several separate network as base classifiers. On the other hand, while these base classifiers are not independent from each other, they show complementary behaviour, based on our experimental findings. More implementation details are discussed in Sec. 3.3.

3 Experimental Evaluation

We evaluated the effectiveness of our approach using the widely used CELEBA and LFWA datasets, described in Section 3.1. Data augmentation techniques

used while training are presented in Section 3.2. In Section 3.3, the network and implementation details are explained. Finally, in Section 3.4, the performance of our proposed method is evaluated along with a comparison with several state-of-the-art techniques.

3.1 Datasets

Our experiments are conducted on two well-known datasets for face attributes classification to assess our proposed method, CELEBA and LFWA [13].

CELEBA [13] consists of $202,599$ images of $10,177$ different celebrity faces identities. The first $8k$ identities are set for training (in total around 160k images), while the remaining images are used for validation and testing (around 20k images each). The dataset provides 5 landmark locations (both eyes, nose, and mouth corners), along with ground-truth for 40 binary attributes for each image.

LFWA [13] is originally constructed for face identification and verification [9], but recently, it is annotated with the same 40 binary attributes. The annotated dataset contains 13,143 images of 5,749 different identities. The dataset has a designated training set portion of 6,263 images, while the rest is reserved for testing. LFWA is one of the challenging datasets with large variations in pose, contrast, illumination and image quality.

3.2 Data Augmentation

Deep networks typically have large number of free parameters on the order of several millions, which makes the networks prone to overfitting. One way to combat overfitting is to use data augmentation. Recently, several advanced methods for face data augmentation have been developed and automated as in [4].

In this work, we want to show the effectiveness of our stand-alone architecture without using sophisticated data augmentation or pre-processing techniques. Therefore, we only use the following simple, but effective data augmentation techniques: (1) Rotation: training images are rotated using a random rotation angle between [-5, +5] around the origin. (2) Scaling: images are scaled up and down with a random scale factor up to a quarter of the image size. (3) Contrast: by converting the color space of the images from RGB to HSV and randomly multiplying the S and V channels with a factor range between [0.5, 1.5]. In addition, blurring with two different filter size (3x3 and 5x5) and histogram equalization are performed.

At every iteration, we randomly decide whether to apply a transformation to the input image and then pick its parameter randomly. Thus, an input image may undergo a combination of multiple transformations, during one presentation.

3.3 Network Details and Implementation

As mentioned in Section 2.3, ResNet-50 is used as our base model in this work, chosen due to its relatively small size and good performance.

All of the layers of ResNet-50 are shared among all of the attributes, up until the last residual block, namely res5b. Then, seven forks are branched from the res5b layer, one for each group of attributes. Each group's shared layers are similar to the layers in the last residual block of ResNet-50, which are as following: a dropout layer followed by a three consecutive blocks of convolutional layer, batch normalization, scaling and ReLU layer.

After every group block, several forks are branched, one for each attribute: a dropout layer, pool layer, followed by a fully connected layer with one unit. The output coming from all of the branches are then concatenated to form a vector of 40 units and a hyperbolic tangent ($tanh$) activation layer is applied after this layer. Finally, a fully connected layer with 40 units is added at the end, followed by $tanh$ activation layer, to learn the correlations among attributes.

For the within-network ensemble, 5 base classifiers are branched after the res2c, res3c, res4a, res4d and res5a layers of the network. The whole network is trained at once, with 7 terms in the loss function (5 coming from the extra branched layers and 2 from the last two fully connected layers).

The implementation is done using the ResNet-50 models provided in the Matlab deep learning toolbox. Throughout this work, we set the batch size equal to 32 and the initial learning rate as 10^{-3} with a total of 20 epochs with stochastic gradient descent for parameters optimization.

The training of the three models effectively took the same amount of time. Specifically, training ResNet-50 model using LFWA dataset for one epoch was performed in 15.52 minutes with the multi-task learning network, 16.02 minutes with the end-to-end network and 18.28 minutes with the within-network-ensemble approach.

3.4 Results and Evaluation

A comparison between our proposed methods that are described in Sec. 2, is shown using the LFWA dataset in Fig. 3. We have obtained an average accuracy of 85.15% using the base system approach; 85.66% with the multi-task network using attribute grouping; 85.92% after embedding an extra layer to capture the relation between the attributes; and finally 86.63% using our novel within-network ensemble technique. Our approach outperforms the state-of-the-art results on LFWA ([6]) by 0.48%.

In Fig. 4, our within-network ensemble approach is compared with the state-of-the-art accuracies obtained on the larger CELEBA dataset. We obtained an average accuracy of 93.20% that surpasses the state-of-the-art obtained in [6], by 0.60%. Note that improvements are small due partly to the already high accuracy rates for this problem and the fact that some of the binary attributes are in fact continuous attributes (e.g. smile).

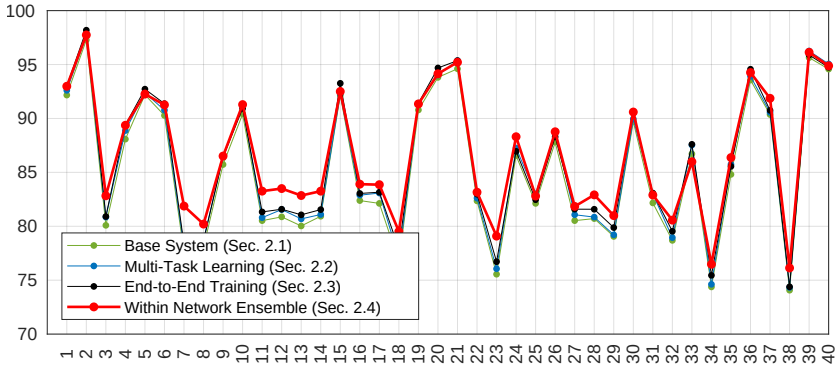

Fig. 3. Obtained accuracies on LFWA dataset from the increasingly complex networks described in Sec. 2. Best viewed in color.

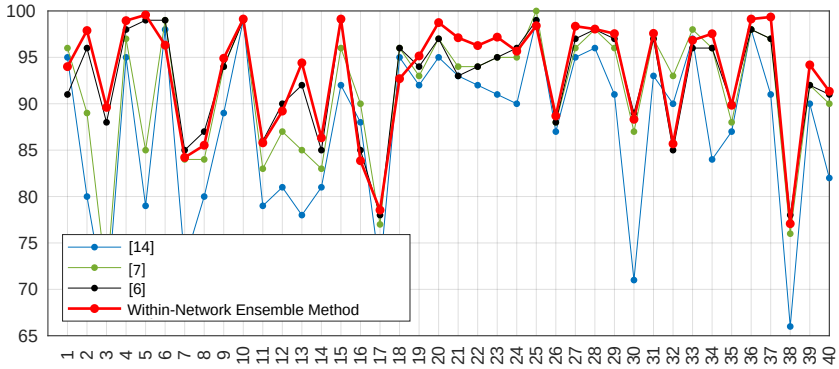

Fig. 4. State-of-the-art accuracies on CELEBA dataset compared with our proposed approach. Best viewed in color.

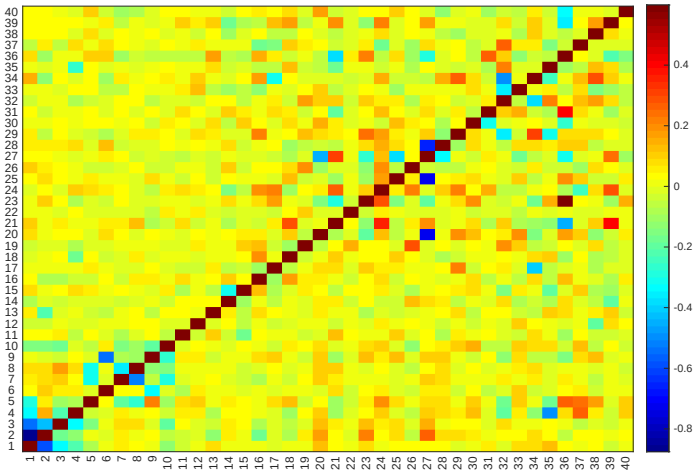

Fig. 5. Learned weights of the last hidden layer that capture the relation between attributes (attributes order is same as in Table 1).

#	Attribute	Baseline	[13]	[7]	[6]	This Work
Head Group						
1	Black Hair	72.84%	95%	**96%**	91%	94.00%
2	Blond Hair	86.67%	80%	89%	96%	**97.89%**
3	Brown Hair	82.03%	68%	71%	88%	**89.61%**
4	Gray Hair	96.81%	95%	97%	98%	**98.96%**
5	Bald	97.88%	79%	85%	99%	**99.57%**
6	Bangs	84.43%	98%	**99%**	**99%**	96.32%
7	Straight Hair	79.01%	73%	84%	**85%**	84.21%
8	Wavy Hair	63.60%	80%	84%	**87%**	85.53%
9	Receding Hairline	91.51%	89%	94%	94%	**94.90%**
10	Wearing Hat	95.80%	99%	99%	99%	**99.13%**
Eyes Group						
11	Arched Eyebrows	71.56%	79%	83%	**86%**	85.79%
12	Narrow Eyes	85.13%	81%	87%	**90%**	89.21%
13	Bushy Eyebrows	87.05%	78%	85%	92%	**94.41%**
14	Bags Under Eyes	79.74%	81%	83%	85%	**86.33%**
15	Eyeglasses	93.54%	92%	96%	99%	**99.13%**
Nose Group						
16	Big Nose	78.80%	88%	**90%**	85%	83.86%
17	Pointy Nose	71.43%	72%	77%	78%	**78.54%**
Mouth Group						
18	Big Lips	67.30%	95%	**96%**	**96%**	92.70%
19	Smiling	49.97%	92%	93%	94%	**95.15%**
20	Mustache	96.13%	95%	97%	97%	**98.75%**
21	Wearing Lipstick	47.81%	93%	94%	93%	**97.11%**
22	Mouth Slightly	50.49%	92%	94%	94%	**96.27%**
Cheek Group						
23	5 o'Clock Shadow	90.01%	91%	95%	95%	**97.18%**
24	Rosy Cheeks	92.83%	90%	95%	**96%**	95.66%
25	Goatee	95.42%	99%	**100%**	99%	98.41%
26	High Cheekbones	51.82%	87%	88%	88%	**88.69%**
27	No Beard	14.63%	95%	96%	97%	**98.36%**
28	Sideburns	95.36%	96%	98%	98%	**98.05%**
Shoulder Group						
29	Double Chin	95.43%	91%	96%	97%	**97.56%**
30	Wearing Necklace	86.21%	71%	87%	**89%**	88.32%
31	Wearing Necktie	92.99%	93%	97%	97%	**97.58%**
General						
32	Attractive	50.42%	90%	**93%**	85%	85.68%
33	Blurry	94.94%	97%	**98%**	96%	96.84%
34	Chubby	94.70%	84%	96%	96%	**97.54%**
35	Young	24.29%	87%	88%	**90%**	89.84%
36	Male	61.35%	98%	98%	98%	**99.13%**
37	Pale Skin	95.79%	91%	97%	97%	**99.35%**
38	Oval Face	70.44%	66%	76%	**78%**	77.07%
39	Heavy Makeup	59.50%	90%	92%	92%	**94.19%**
40	Wearing Earrings	79.34%	82%	90%	91%	**91.34%**
	Average	76.87%	87.30%	91.32%	92.60%	**93.20%**

Table 2. State-of-the-art accuracies on CELEBA dataset compared with the results obtained in this work, using the within-network ensemble. Bold figures indicate the best results.

By visualizing the learned weights of the last hidden layer (Fig. 5), we found that the relationship between attributes are nicely captured. For instance, the learned weights show a high negative correlation between "No Beard" attribute and "Mustache", "Goatee", and "Side Burns" attributes. Contrarily, there is a high positive correlation between "Heavy Makeup" attribute and "Wearing Lipstick", "Rosy Cheeks", and "No Beard" attributes.

State-of-art results on the CELEBA dataset and those obtained with the within-network ensemble are shown in Table 2.

4 Conclusion

We present an end-to-end multi-task framework for face attribute classification that considers attribute location to reduce network size and correlation among attributes to improve accuracy.

We also introduce a novel ensemble technique that we call within-network ensemble, by branching output nodes from different depths of the network and computing the loss over all these branches. As the network is shared, this branching results in very little computational overhead. To the best of our knowledge, this ensemble technique has not been suggested before, while it brings non-negligible improvements (0.71% points accuracy improvement over the end-to-end network). Our results surpass state-of-the-art on both LFWA and CELEBA datasets, with 86.63% and 93.20% average accuracies, respectively.

5 Acknowledgements

We gratefully acknowledge NVIDIA Corporation with the donation of the Titan X Pascal GPU used in this research.

References

1. Aly, S.A., Yanikoglu, B.: Multi-label networks for face attributes classification. In: IEEE International Conference on Multimedia & Expo Workshops (ICMEW). pp. 1–6. IEEE (2018)
2. Baxter, J.: A bayesian/information theoretic model of learning to learn via multiple task sampling. Machine Learning **28**(1), 7–39 (1997)
3. Bourdev, L., Maji, S., Malik, J.: Describing people: A poselet-based approach to attribute classification. In: International Conference on Computer Vision (ICCV). pp. 1543–1550. IEEE (2011)
4. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation policies from data. arXiv:1805.09501 (2018)
5. Ehrlich, M., Shields, T.J., Almaev, T., Amer, M.R.: Facial attributes classification using multi-task representation learning. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 47–55 (2016)
6. Han, H., Jain, A.K., Wang, F., Shan, S., Chen, X.: Heterogeneous face attribute estimation: A deep multi-task learning approach. IEEE transactions on Pattern Analysis and Machine Intelligence **40**(11), 2597–2609 (2018)

7. Hand, E.M., Chellappa, R.: Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification. In: 31st AAAI Conference on Artificial Intelligence (2017)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
9. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. rep., 07-49, University of Massachusetts, Amherst, Technical Report (October, 2007)
10. Huang, W., Song, G., Hong, H., Xie, K.: Deep architecture for traffic flow prediction: deep belief networks with multitask learning. IEEE Transactions on Intelligent Transportation Systems **15**(5), 2191–2201 (2014)
11. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: IEEE 12th International Conference on Computer Vision (ICCV). pp. 365–372. IEEE (2009)
12. Li, Y., Wang, R., Liu, H., Jiang, H., Shan, S., Chen, X.: Two birds, one stone: Jointly learning binary code for large-scale face image retrieval and attributes prediction. In: IEEE International Conference on Computer Vision (ICCV). pp. 3819–3827 (2015)
13. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: IEEE International Conference on Computer Vision (ICCV). pp. 3730–3738 (2015)
14. Luo, Y., Tao, D., Geng, B., Xu, C., Maybank, S.J.: Manifold regularized multitask learning for semi-supervised multilabel image classification. IEEE Transactions on Image Processing **22**(2), 523–536 (2013)
15. Ranjan, R., Patel, V.M., Chellappa, R.: Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. arXiv:1603.01249 (2016)
16. Rozsa, A., Günther, M., Rudd, E.M., Boult, T.E.: Are facial attributes adversarially robust? In: 23rd International Conference on Pattern Recognition (ICPR). pp. 3121–3127. IEEE (2016)
17. Rudd, E.M., Günther, M., Boult, T.E.: Moon: A mixed objective optimization network for the recognition of facial attributes. In: European Conference on Computer Vision (ECCV). pp. 19–35. Springer (2016)
18. Sharif Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: an astounding baseline for recognition. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 806–813 (2014)
19. Song, F., Tan, X., Chen, S.: Exploiting relationship between attributes for improved face verification. Computer Vision and Image Understanding **122**, 143–154 (2014)
20. Yi, S., Jiang, N., Feng, B., Wang, X., Liu, W.: Online similarity learning for visual tracking. Information Sciences **364**, 33–50 (2016)
21. Zhong, Y., Sullivan, J., Li, H.: Face attribute prediction using off-the-shelf CNN features. In: International Conference on Biometrics (ICB). pp. 1–7. IEEE (2016)
22. Zhu, Z., Luo, P., Wang, X., Tang, X.: Multi-view perceptron: A deep model for learning face identity and view representations. In: Advances in Neural Information Processing Systems (NIPS). pp. 217–225 (2014)