Deep Facial Attribute Detection in the Wild: From General to Specific

Yuechuan Sun
ycsun@mail.ustc.edu.cn
Jun Yu
harryjun@ustc.edu.cn

Department of Automation
University of Science and Technology of China
Hefei, China

Abstract

Accurate facial attribute interpretation is a challenging task in real life due to large head poses, occlusion and illumination variations. This work proposes a general-to-specific deep convolutional network architecture for predicting multiple attributes from a single image in the wild. First, we model the interdependencies of local facial regions by joint learning of all the attributes. Second, task-aware learning is established to explore the disparity regarding each attribute. Finally, an attribute-aware face cropping scheme is proposed to extract more discriminative features from where a certain attribute naturally shows up. The proposed general-to-specific learning strategy ensures both robustness and performance of our model. Extensive experiments on the CelebA and LFWA datasets demonstrate the effectiveness of our architecture and the superiority to state-of-the-art alternatives.

1 Introduction

The problem of analyzing facial attributes (e.g., gender, hairstyle, smile) using computer vision techniques attracts extensive research interest due to its potential real-life applications in surveillance[18], entertainment[12], medical treatment[7], etc. Accurate facial attribute detection also benefits a number of facial analysis tasks, such as face verification[10, 16], retrieval[9, 11] and alignment[24]. However, predicting facial attributes in real-world scenarios still remains challenging because faces vary dramatically under different poses, occlusions, and lighting conditions. For example, eyes are hardly visible when the identity wears glasses, which leads to the intense difficulty of eye shape analysis.

Existing method addressing attribute recognition can be generally categorized into global[13, 21] and local[1, 2, 10, 14, 23] ones. Global methods usually extract features from the entire face and do not require localization of landmarks or object parts. They assume that different attributes are interdependent and each face part should be equally considered. All attributes are treated equivalently and no customized processing is conducted. On the contrary, local methods treat each attribute independently by first detecting face parts and applying feature descriptors to each part for training a classifier, where face alignment is of vital importance to the final result. Local methods generally outperform the global ones when reliable preprocessing is available as distinct spatial features associated to each attribute is captured and less extra noise information is introduced. However, they may fail under unconstrained condition where accurate face localization and alignment are difficult to obtain and faces are partially visible for many reasons. Considering the pros and cons of two methods mentioned before, our work integrates both methods by extracting facial representations from the holistic region of the aligned face and emphasizing on the functional parts in which a certain attribute naturally shows up using different cropping schemes.

Motivated by the recent success of deep convolutional neural networks (CNNs) on facial attribute analysis [6, 13, 16, 19], we propose a CNN based approach for facial attribute prediction under unconstrained conditions. By introducing the general-to-specific learning strategy, we implicitly discover the correlations of all the attributes considered while specifically focus on the distinctions. The key contributions of this paper are:

1. Traditionally, all attributes are disconnected and treated equally [9, 13], we overcome this limitation by proposing a general-to-specific learning framework that extracts both interconnections and disparities to improve facial attribute detection under uncontrolled conditions. Through multi-task learning, our model yields higher robustness to challenging scenarios by learning interdependencies of different attributes. Meanwhile, distinct information are captured in separate learning using task-aware face cropping and used to ensure exceptional performance.

2. We show superior attribute prediction performance over the state-of-the-art methods[6, 9, 13, 16, 19] on the biggest public benchmark dataset for facial attribute analysis, i.e., CelebA dataset[13]. We also outperform all the other methods which do not introduce external datasets on LFWA[5, 13] dataset.

The rest of this paper is organized as follows. The details of the proposed general-to-specific attribute detection architecture are described in Section 2. Experiments on two datasets are reported in Section 3. Section 4 concludes the paper.

2 Proposed Approach

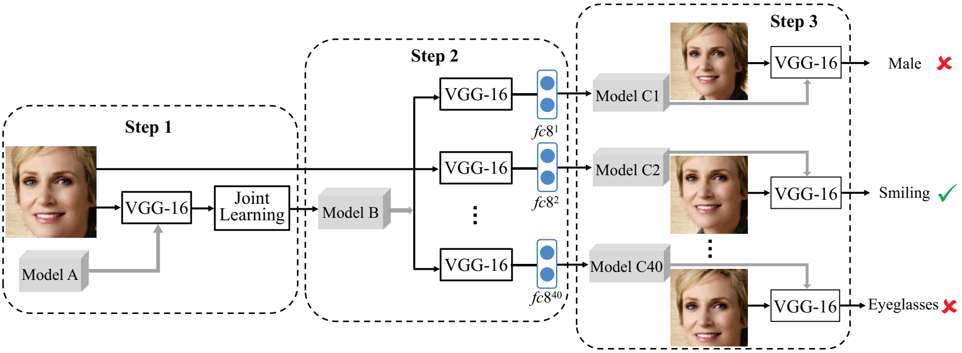

Figure 1: Overview of the training procedure. The learning incorporates three steps and features are refined step-to-step. Note that only Step 3 is required for attribute detection of a new testing image.

Fig. 1 shows the framework of the proposed method, which consists of three continuous parts: the joint learning in Step 1, the task-aware learning in Step 2 and the attribute-aware face cropping in Step 3. In Step 1, the face images are cropped and fed into the VGG-16 network for jointly learning all attributes, and the network weights are initialized with Model A (the pre-trained VGG-Face weight model). When Step 1 is finished, we can obtain the Model B, which is then used for initializing the weights of 40 separate networks in Step 2 to learn each attribute by optimizing the corresponding binary classification problem. In Step 3, the face images are distinctively cropped according to the physiological property of each attribute. Networks are initialized with the weights (Model C) obtained in Step 2 and trained using the cropped images. The attribute classification result of the networks in Step 3 is considered as the final output.

2.1 Network Architecture

Our CNN architecture follows the design of VGG-16[17], which was successfully used in image classification[17], face recognition[15] and so on. The network consists of 13 convolutional layers, 5 pooling layers and 3 fully connected layers.All hidden layers are equipped with the rectification (ReLU) non-linearity[8]. The final output of the network is fed to a soft-max classifier to produce class probabilities. All the convolutional layers and the first two fully connected layers are exactly the same as the corresponding layers in[17]. However, the last fully connected layer (*fc8*) has only two neurons as we aim to make a binary judgement on the presence of a given attribute. Specifically, the network in Step 1 has 80 outputs (two for each attribute) while those in Step 2 and Step 3 have only two. The input to the network is an RGB face image cropped and scaled to 224×224 pixels. The network is initialized with the *VGG-Face* model[15] (refer to the Weight A in Fig. 1), which was pre-trained on a large-scale face recognition dataset created by [15]. We train the network using cross-entropy loss. Specifically, Section 3.1 provides further details of the training procedure.

2.2 General to Specific Learning

Considering the inherent physiological structure of human face, we believe that all the attributes to be detected yield shared interconnections as well as unique disparities. Therefore, it is straightforward to utilize the potent correlations for model robustness through joint learning and boost individual attribute detection performance via separate learning.

Multi-task Joint Learning. Multi-task learning seeks to explore the related tasks on the same data and maximize the overall accuracy over all tasks. By joint learning, we can discover the underlying interrelationship among these tasks, which is unobtainable from any single task. When large face variations are present, the separate models for each attribute become unreliable due to feature corruption at the corresponding face regions. On the other hand, facial attribute are highly connected and the presence of some attributes implies the presence or absence of others. For instance, there is a great possibility that a person with *mustache* is *male*, whereas the presence of *lipstick* is hardly possible at the meantime. Therefore, joint learning of all the attributes not only benefits feature extraction for the detection of each attribute, but also contributes to the exploration of implicit correlations among them (either positive or negative). Features derived from various attributes have higher robustness and generalization in that they incorporate general information which ameliorates the detection of each attribute.

Therefore, a loss function, which mixes all attribute predictions and performs simultaneous optimization, should be designed. Given N training images $\{x_n\}_{n=1}^{N}$ and their corresponding binary labels of T attributes $\{y_n^t\}_{n=1,t=1}^{N,T}$, the objective of learning the T tasks is to minimize the overall loss as:

$$L_J(W) = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T} L_n^t(W), \tag{1}$$

where $L_n^t(W)$ is the cross-entropy loss of the t-th attribute on the n-th image and is parameterized by a weight matrix W. Here, we treat each attribute equally. The individual loss is formulated as:

$$L_n^t(W) = -y_n^t \log p_n^t - (1-y_n^t)\log(1-p_n^t), \tag{2}$$

here p_n^t denotes the probability of the presence of the t-th attribute on the n-th image.

Task-aware Learning. Despite of the high universality of the feature obtained from joint learning, personalized learning is still required for exploring the distinction of each attribute. In the previous section, we treat all attributes equally (refer to Eq.1). However, this is not reasonable considering the dominance of some attributes. For instance, *male* influences the presence of various gender-related attributes and should be given more emphasis. To address this problem, it is necessary to apply task-aware learning to focus on the target attribute and remove the adverse effect resulting from others.

In this step, individual networks for each attribute are trained using the loss given in Eq.2. Specifically, we fine-tune the networks using the network weight obtained in Step 1 (Weight B) for feature inheritance, and the same image data (see Fig. 2(a)) is used as input.

Attribute-aware Face Cropping. The aforementioned training procedures were all conducted using the same cropping scheme (see Fig. 2(a)), which is not tailored for each attribute. However, facial attributes generally originate from corresponding facial organs or parts. Irrelevant face regions may lead to feature disturbance more or less. Hence, applying spatial attention can wipe out the adverse impact posed by trivial face areas and encourage the networks to extract more customized information accordingly. For example, we can easily distinguish smile from lower face and forehead is thus redundant whereas the entire face region is demanded for gender estimation. To this end, we propose using different cropping schemes to eliminate spatial irrelevancy and emphasize on the corresponding face regions by changing *margin* M and *offset* O.

Given the detected rectangle face region $\{X,Y,W,H\}$ after alignment using MTCNN[22], where (X,Y) is the coordinate of the top left corner and W, H denote the weight and height of the face. Suppose that H is greater than W, the personalized cropped face determined by M and O can be formulated as:

$$\{X - \frac{(M+1)\times H - W}{2}, Y + \frac{(2\times O - M)\times H}{2}, (1+M)\times H, (1+M)\times H\}. \tag{3}$$

By varying M and O, we can obtain face regions with specified emphasis. In this work, altogether 8 types of cropping schemes are selected as shown in Fig. 2.

Similar to Step 2, each attibute in Step 3 is leared by fine-tuning the network with weight obtained in the previous step (Weight $C_i, i = 1,2,...,40$). Once Step 3 is done, the entire learning process finishes and we can obtain the final network weights that can be used for classifying attributes of new face images.

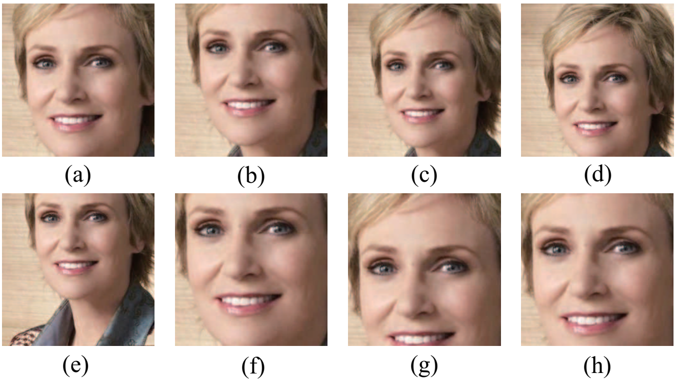

Figure 2: Examples of all the task-aware face cropping schemes used. M and O for each scheme are: (a) 0, 0, (b) 0, 0.1, (c) 0.1, 0, (d) 0.1, -0.1, (e) 0.1, 0.5, (f) -0.1, 0, (g) -0.1, -0.1, (h) -0.1, 0.1.

3 Experiments

We evaluate our proposed approach on two datasets (i.e., CelebA and LFWA) and compare with state-of-the-art architectures. A workstation with Intel i7-7700K 4.2G, 32G memory and NVIDIA GTX1080 Ti is used for the experiments.

CelebA. As the largest facial attribute dataset available, the CelebA dataset consists of over 200k images from approximately 10k celebrities. Following the standard evaluation protocol, the first 160k images are used for training, 20k images for validation and the remaining 20k for testing. Each image is annotated with binary labels of 40 face attributes, as well as 5 key points (both eyes, the mouth corners and the nose tip). In addition to the original images, CelebA provides a set of pre-cropped images with alignment. The raw images are used for our experiments.

LFWA. LFWA is created based on the LFW face dataset[5]. It contains a total of 13,143 images of 5,749 identities with training and testing splits which divide the dataset into two roughly equal partitions. Each image is labeled with the same 40 attributes as in CelebA.

3.1 Implementation Details

The publicly available face detector MTCNN[22] is used for face detection and landmark localization. Face images are first rotated according to the eye coordinates. When the detection fails, we discard the image if it is in training or validation set, but use the provided face regions and landmarks if it is a testing image. Facial ROI is then obtained by attribute-aware face cropping and scaled to 240×240. A 224×224 random patch as well as its horizontal reflection is extracted for network training. The 224×224 image is directly used for testing without argumentation, which fastens the running time considerably at the cost of classification accuracy compared with [8] and [17]. Note that only Step 3 is required during testing. In opposition to [16] that uses aligned images provided by CelebA, we directly use the raw face images, which are much more challenging and meaningful.

The Caffe deep learning toolbox [23] is used for the implementation of our network. We employ the stochastic gradient descent to train the networks with a batch size of 128, momentum of 0.9, and weight decay of 0.0005. The base learning rate is 0.0001 and decreased by inverse decay with $\gamma = 0.0001$ and power $= 0.5$. According to [20], bottom layers contain more generic features and more feature transition should be made by top layers when fine-tuning. Hence, we set *lr_mult* as 1 on all convolutional layers while we set it as 3, 5 and 8 on the last three fully connected layers. We run the trainings for a total of 120k and 40k iterations in Step 1 and Step 2, 3 respectively.

3.2 Effectiveness of the Proposed Framework

We evaluate the effectiveness of the proposed general-to-specific learning strategy shown in Fig. 1 by combining different learning steps, and the results of all of possible combinations are reported in Table 1. The last row represents the proposed learning strategy which involves all three steps. It can be observed that the proposed learning scheme works well on both datasets. Joint learning (Step 1) alone achieves decent performance by learning correlations which apply to each attribute. Taking advantage of explicit and implicit relationships among attributes allows for better feature representation that can contribute to final attribute classification. Step 2 and 3 can further enhance feature representation ability by extracting distinctive information. If we combine either two steps, the model performance can be promoted. Applying the complete learning process yields the highest attribute prediction accuracies. It is notable that our method demonstrates higher performance improvement while lower average accuracy on LFWA than on CelebA. This is likely due to the lack of training data in LFWA and thus overfitting may happen when the model is trained on a single task. Hence, our method is particularly suitable for small-scale dataset. Comparing different learning methods, we can conclude that either stage is indispensable for effective attribute prediction.

Table 1: Comparison of different training stages on average attribute prediction accuracy.

Method	CelebA(%)	LFWA(%)
Step 1	90.6	84.9
Step 2	90.9	83.2
Step 3	91.2	85.3
Step 1,2	91.1	85.2
Step 1,3	91.4	86.6
Step 2,3	91.2	86.3
Step 1,2,3	**91.6**	**87.1**

3.3 Comparison with Benchmark Methods

Competitors. In this section, we compare our method with state-of-the-art attribute prediction ones in terms of classification accuracy on the abovementioned two datasets. FaceTracer[9] extracts hand-crafted features (HOG and color histogram) of different functional face regions to train an SVM classifier for attribute classification. LNets+ANet[13] cascades two localization networks (LNets) and an attribute prediction network (ANet) to automatically detect the face region and learn the facial attributes from the detected part accordingly. PANDA [23] leverages CNNs to extract attribute features related to body pose. LNets+ANet [12] uses a total of three networks to automatically detect the face region and learn the facial attributes from the detected part accordingly. Moon[16] applies a mixed objective optimization network to learn all attributes. Walk-and-Learn[19] collects an egocentric video dataset with weather and location context to facilitate attribute learning. AFFACT[4] proposes an alignment-free data augmentation technique using an ensemble of three ResNets for attribute classification. Kalayeh *et al.* [6] employ semantic segmentation to improve facial attribute prediction and achieve the state-of-the-art performance on both datasets.

Table 2: Performance comparison of attribute detection with state-of-the-art methods on CelebA.

	FaceTracer[9]	LNets+ANet[13]	Moon[16]	Walk-and-Learn[19]	AFFACT[4]	Kalayeh *et al.* [6]	Ours
5ClockShadow	85	91	94	84	**94.8**	94.5	94.7
Arch. Eyebrows	76	79	82.3	**87**	83.9	83.1	83.6
Attractive	78	81	81.7	**84**	82.9	82.3	83.2
Bags Under Eyes	76	79	84.9	**87**	85.2	85.4	85.5
Bald	89	98	98.8	92	**99.1**	98.8	99.0
Bangs	88	95	95.8	96	96.1	95.5	**96.2**
Big Lips	64	68	71.5	**78**	72.5	71.7	71.5
Big Nose	74	78	84	**91**	84.4	84.5	85.0
Black Hair	70	88	89.4	84	**90.5**	90.1	90.2
Blond Hair	80	95	95.9	92	**96.2**	95.8	96.1
Blurry	81	84	95.7	91	96.0	95.7	**96.4**
Brown Hair	60	80	**89.4**	81	88.5	89.2	89.0
Bushy Eyebrows	80	90	92.6	**93**	92.3	92.4	**93.0**
Chubby	86	91	95.4	89	95.7	95.6	**95.9**
Double Chin	88	92	96.3	93	**96.4**	96.3	**96.4**
Eyeglasses	98	99	99.5	97	99.6	99.3	**99.7**
Goatee	93	95	97	92	97.5	97.3	**97.6**
Gray Hair	90	97	98.1	95	**98.3**	98.2	98.2
Heavy Makeup	85	90	91	**96**	92.0	90.8	91.8
High Cheekbones	84	87	87	**95**	87.6	87.1	88.1
Male	91	98	98.1	96	98.2	97.7	**98.8**
Mouth S. O.	87	92	93.5	**97**	93.8	92.2	94.1
Mustache	91	95	96.8	90	**97.0**	**97.0**	96.9
Narrow Eyes	82	81	86.5	79	87.6	86.7	**87.7**
No Beard	90	95	95.6	90	96.2	95.7	**96.2**
Oval Face	64	66	75.7	**79**	76.6	77.8	74.8
Pale Skin	83	91	97	85	**97.1**	**97.1**	**97.1**
Pointy Nose	68	72	76.5	77	77.1	76.5	**77.8**
Receding Hairline	76	89	93.6	84	93.7	93.3	**93.9**
Rosy Cheeks	84	90	94.8	**96**	95.2	94.8	95.1
Sideburns	94	96	97.6	92	97.8	97.7	**98.0**
Smiling	89	92	92.6	**98**	92.8	91.9	93.3
Straight Hair	63	73	82.3	75	**85.0**	83.6	83.8
Wavy Hair	73	80	82.5	85	**85.7**	84.8	84.3
Wearing Earrings	73	82	89.6	**91**	**91.0**	90	90.5
Wearing Hat	89	99	99	96	99.1	98.8	**99.2**
Wearing Lipstick	89	93	93.9	92	93.7	93.6	**94.3**
Wearing Necklace	68	71	87	77	88.3	88.7	**89.3**
Wearing Necktie	86	93	96.6	84	96.9	97.1	**97.3**
Young	80	87	88.1	86	**88.9**	87.8	**88.9**
Average	81.1	87.3	90.9	88.7	91.5	91.2	**91.6**

Evaluation on CelebA. As shown in Table 2, our method attains the highest average accuracy of 91.6% on CelebA, significantly improving on FaceTracer[9] by 13%. Despite the various challenges caused by unconstrained conditions, we still obtain remarkable performance and accurately classify over 99% of the images on some attributes, i.e., *eyeglasses* and *wearing hats*. All compared methods except the method in [16] disconnect face attributes and learn each one independently, which leads to the loss of beneficial interrelatedness. Moon [16]incorporates multi-task learning, however, the further task-aware learning is ignored. As for the LFWA dataset, our approach outperforms all compared ones that do not rely on external datasets. It should be noted that the method in [6] uses a semantic dataset for extracting prior information, while we still outperform them on 33 attributes and yield higher average accuracies.

Table 3: Performance comparison of attribute detection with state-of-the-art methods on LFWA.

	FaceTracer[9]	PANDA[23]	LNets+ANet[13]	Walk-and-Learn[19]	Kalayeh *et al.* [6]	Ours
5ClockShadow	70	**84**	**84**	76	83.7	78.4
Arch. Eyebrows	67	79	82	82	80.9	**83.9**
Attractive	71	81	83	82	**85.1**	80.4
Bags Under Eyes	65	80	83	91	**92.8**	84.9
Bald	77	84	88	82	91.8	**92.5**
Bangs	72	84	88	**93**	80.2	91.5
Big Lips	68	73	75	75	**84.7**	80.6
Big Nose	73	79	81	92	**92.8**	84.9
Black Hair	76	87	90	93	**97.7**	91.9
Blond Hair	88	94	97	97	87.5	**97.5**
Blurry	73	74	74	86	82.7	**87.3**
Brown Hair	62	74	77	83	**85.8**	80.9
Bushy Eyebrows	67	79	82	78	77.7	**87**
Chubby	67	69	73	79	**81.9**	76.6
Double Chin	70	75	78	81	**92.8**	83.3
Eyeglasses	90	89	**95**	94	84.1	92.6
Goatee	69	75	78	80	**89.2**	84.5
Gray Hair	78	81	84	91	**95.9**	88.8
Heavy Makeup	88	93	95	96	89.5	**96.2**
High Cheekbones	77	86	88	**96**	94.4	89.7
Male	84	92	94	93	94.4	**96.1**
Mouth S. O.	77	78	82	**94**	84.3	83.7
Mustache	83	87	92	83	94	**94.3**
Narrow Eyes	73	73	81	79	84.7	**85.1**
No Beard	69	75	79	75	**83.6**	82
Oval Face	66	72	74	**84**	77.9	79.5
Pale Skin	70	84	84	87	**91.1**	89.1
Pointy Nose	74	76	80	**93**	85	85.4
Receding Hairline	63	84	85	86	86.6	**86.7**
Rosy Cheeks	70	73	78	81	**86.3**	85.9
Sideburns	71	76	77	77	83.2	**83.8**
Smiling	78	89	91	**97**	92.5	92.5
Straight Hair	67	73	76	76	81.6	**81.9**
Wavy Hair	62	75	76	**89**	81.2	81.6
Wearing Earrings	88	92	94	**96**	95.2	95.1
Wearing Hat	75	82	88	86	**91.1**	90.6
Wearing Lipstick	87	93	95	**97**	95.2	95.4
Wearing Necklace	81	86	88	**95**	90.1	90.2
Wearing Necktie	71	79	79	80	**83.9**	83.3
Young	80	82	86	**89**	86.9	86.7
Average	73.9	81	83.9	86.6	**87.1**	**87.1**

Evaluation on LFWA. Table 3 shows the results for the LFWA dataset. We can see that our method achieves comparable average accuracy of 87.1% to Kalayeh *et al.* [6] and outperforms all compared ones that do not rely on external datasets. Considering that LFWA is a small-scale dataset, the effect of introducing external datasets for prior information becomes more obvious. Hence, our method is much more practical and efficient in real-life applications as it does not require any additional data or processing. The effectiveness of the proposed general-to-specific learning is fully verified again by comparing to previous works. As mentioned before, all methods present inferior performance on LFWA than on CelebA, which is partly due to the model overfitting.

4 CONCLUSION

This paper proposed a deep learning based framework for robust facial attribute prediction in the wild. We present a general-to-specific learning strategy composed of three stages to obtain step-to-step optimization. Different from previous approaches, our method leverages attribute correlations to retain high robustness and generalizability via multi-task joint leaning. Besides, the proposed task-aware learning ensures the distinction of each attribute and attribute-aware cropping scheme minimizes adverse information from irrelevant face parts. We evaluate our approach on CelebA and LFWA datasets and achieve state-of-the-art performance. For future work we plan to introduce attention model [3] for automatically localizing discriminative regions.

5 ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China (U1736123, 61572450, 61303150), Anhui Provincial Natural Science Foundation (1708085QF138), the Fundamental Research Funds for the Central Universities (WK2350000002).

References

[1] Thomas Berg and Peter N Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *CVPR*, pages 955–962, 2013.

[2] Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Describing people: A poselet-based approach to attribute classification. In *ICCV*, pages 1543–1550. IEEE, 2011.

[3] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *CVPR*, pages 4438–4446, 2017.

[4] Manuel Günther, Andras Rozsa, and Terrance E Boult. Affact-alignment free facial attribute classification technique. In *IJCB*, 2017.

[5] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[6] Mahdi M Kalayeh, Boqing Gong, and Mubarak Shah. Improving facial attribute prediction using semantic segmentation. *CVPR*, 2017.

[7] Sebastian Kaltwang, Ognjen Rudovic, and Maja Pantic. Continuous pain intensity estimation from facial expressions. *Advances in Visual Computing*, pages 368–377, 2012.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[9] Neeraj Kumar, Peter Belhumeur, and Shree Nayar. Facetracer: A search engine for large collections of images with faces. In *ECCV*, pages 340–353. Springer, 2008.

[10] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Attribute and simile classifiers for face verification. In *ICCV*, pages 365–372. IEEE, 2009.

[11] Neeraj Kumar, Alexander Berg, Peter N Belhumeur, and Shree Nayar. Describable visual attributes for face verification and image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1962–1977, 2011.

[12] Yu-Heng Lei, Yan-Ying Chen, Lime Iida, Bor-Chun Chen, Hsiao-Hang Su, and Winston H Hsu. Photo search by face positions and facial attributes on touch devices. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 651–654. ACM, 2011.

[13] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015.

[14] Domingo Mery and Kevin Bowyer. Automatic facial attribute analysis via adaptive sparse representation of random patches. *Pattern Recognition Letters*, 68:260–269, 2015.

[15] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *BMVC*, volume 1, page 6, 2015.

[16] Ethan M Rudd, Manuel Günther, and Terrance E Boult. Moon: A mixed objective optimization network for the recognition of facial attributes. In *ECCV*, pages 19–35. Springer, 2016.

[17] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *in ICLR*, 2015.

[18] Daniel A Vaquero, Rogerio S Feris, Duan Tran, Lisa Brown, Arun Hampapur, and Matthew Turk. Attribute-based people search in surveillance environments. In *WACV*, pages 1–8. IEEE, 2009.

[19] Jing Wang, Yu Cheng, and Rogerio Schmidt Feris. Walk and learn: Facial attribute representation learning from egocentric video and contextual data. In *CVPR*, pages 2295–2304, 2016.

[20] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *NIPS*, pages 3320–3328, 2014.

[21] Kaipeng Zhang, Lianzhi Tan, Zhifeng Li, and Yu Qiao. Gender and smile classification using deep convolutional neural networks. In *CVPR*, pages 34–38, 2016.

[22] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[23] Ning Zhang, Manohar Paluri, Marc'Aurelio Ranzato, Trevor Darrell, and Lubomir Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *CVPR*, pages 1637–1644, 2014.

[24] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, pages 94–108. Springer, 2014.