Journal of
Electronic Imaging

JElectronicImaging.org

Attribute-correlated local regions for deep relative attributes learning

Fen Zhang
Xiangwei Kong
Ze Jia

SPIE. IS&T imaging.org

Fen Zhang, Xiangwei Kong, Ze Jia, "Attribute-correlated local regions for deep relative attributes learning," *J. Electron. Imaging* **27**(4), 043021 (2018), doi: 10.1117/1.JEI.27.4.043021.

Attribute-correlated local regions for deep relative attributes learning

Fen Zhang,[a] Xiangwei Kong,[a,*] and Ze Jia[b]
[a]Dalian University of Technology, School of Information and Communication Engineering, Dalian, China
[b]Unit 91439 of PLA, Dalian, China

Abstract. Relative attributes have a more detailed and accurate description than previous binary ones. We propose to utilize the acquired attribute-correlated local regions of image for learning deep relative attributes. Different from previous works, which usually discover the spatial extent of the corresponding attribute based on the ranking list of all the images in the image set, we first classify the images according to the presence or absence of each provided attribute. Then, we sort the images in the classified image sets using a semisupervised method and learn the most relevant regions corresponding to a specific attribute. The learned local regions in two classified image sets are integrated to obtain the final result. The images and localized regions are then fed into the pretrained convolutional neural network model for feature extraction. Therefore, the concatenation of the high-level global feature and intermediate local feature is adopted to predict the relative attributes. We show that the proposed method produces a competitive performance compared with the state of the art in relative attribute prediction on three public benchmarks. [DOI: 10.1117/1.JEI.27.4.043021]

Keywords: relative attributes; convolutional neural network; local regions; spatial extent; intermediate features.

Paper 180239 received Mar. 16, 2018; accepted for publication Jul. 5, 2018; published online Jul. 25, 2018.

1 Introduction

As intermediate semantic representations, attributes are often adopted in the computer vision community, e.g., fine-grained recognition,[1,2] object classification,[3,4] face verification,[5,6] and image retrieval.[7–9] The main idea is to learn classifiers to predict the presence of various high-level semantic concepts from objects, locations, and activity types. Early works based on the attributes mostly relied on handcrafted features,[10] e.g., SIFT, HOG, and color histogram; however, the performance was limited by the discriminative ability of the low-level handcrafted features.

Recently, the convolutional neural network (CNN)-based deep learning method has been employed as a strong feature learning strategy extensively in some works,[11–18] due to the higher discriminative learning ability. Such a network learns a hierarchy of nonlinear features automatically, which could predict the image attributes[19–23] successfully and achieve attribute-related applications, e.g., face recognition,[24] scene understanding,[25] and clothing retrieval;[26] however, these works mentioned above focus on generating discriminative binary attributes.

For many visual attributes, it is difficult to describe the exact degrees of their presences, whereas the relative ordering of presence can be easily figured out. As opposed to predicting the presence of an attribute, a relative attribute indicates the strength of an attribute in an image, and the relative descriptions are more precise and informative than the binary ones. Some representative relative attributes-based works have been proposed, from which Parikh and Grauman[27] designed more complex and task-specific models on their seminal work; however, the handcrafted visual features[28–34] are employed. Recently, the deep feature representations learned from CNN-based models have been exploited to predict relative attributes.[35,36] For example, Yaser et al.[35] introduced a CNN-based model, which is composed of a feature learning and extracting part and a ranking part for the task of relative attribute prediction. The learned deep feature representations are only global ones based on the whole images. Krishna and Yong[36] proposed an end-to-end deep convolutional network to localize and rank relative visual attributes simultaneously, given only weakly supervised pairwise image comparisons. Motivated by jointly learning the attribute's features, localization, and ranker, this method can achieve a higher performance; however, the training data and effort requirements of this method seem enormous.

Moreover, local representations often lead to better performance compared with global representations in recent work because many attributes are locally orientated.[1,2,5,30,37] For example, the attribute "smile" can be more effectively and easily learned when people's mouth is localized. Therefore, in this paper, we tend to learn relative attributes using a pipeline that is composed of conventional regions localization module, deep feature extraction module, and ranking module. The pipeline is shown in Fig. 1. We focus on discovering the local regions that most relevant to the attributes, and learning proper deep feature representations from a pre-trained CNN model to enhance relative attributes prediction.

To localize the relevant attribute regions, some early work uses pretrained part detectors;[2,30] however, because the part detectors are trained independently of the attribute, the learned parts may not be useful necessarily for modeling the desired attribute. Furthermore, some abstract attributes

*Address all correspondence to: Xiangwei Kong, E-mail: kongxw@dlut.edu.cn

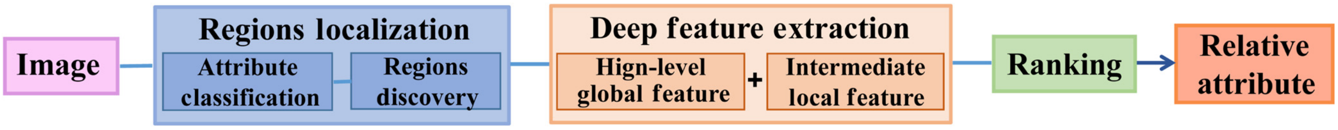

Fig. 1 The pipeline of relative attributes learning, which is composed of regions localization module, deep feature extraction module, and ranking module.

(e.g., good looking) do not have well-defined parts, which mean that modeling a "good looking" detector can be difficult. To address these issues, Xiao and Lee[38] proposed a method that discovers the spatial extent of relative attributes in images across varying attribute strengths automatically, given only weakly supervised pairwise comparisons. The main idea is to generate visual chains along the attribute spectrum, and then select the most relevant ones corresponding to the provided relative attribute annotations. However, since the images are sorted in the entire image set when initializing a single chain for an attribute, the attribute appearance may change not so smoothly among some adjacent images.

Based on the above considerations, in this paper, we propose to roughly classify the images in the entire image set according to the presence or absence of each attribute before discovering the spatial extent of the attributes. This operation could improve the accuracy of the visual chains generation to some extent because the attribute appearance changes more smoothly in each categorized image set. Moreover, inspired by Ref. 19 that the different layers of deep features encode different levels of visual information, we expect that the local CNN features of the localized regions could describe the appearance variations in the corresponding attributes effectively. To this end, the final deep representations for the attributes are formulated by the concatenation of the intermediate local CNN features and the high-level global CNN features, which serve as the inputs of the ranking module.

To verify the effectiveness of the proposed method, we conducted extensive experiments on three public benchmarks: LFW-10, Zappos50K-1, and Shoes. The experimental results show that the proposed method produces a competitive performance compared with the state of the art in relative attribute prediction.

There are three contributions in this paper: (1) an attribute classification procedure is performed rather than directly discovering the spatial extents corresponding to each provided attribute in each image, (2) a semisupervised group sparse-based method is used to sort the images in the classified image sets, as the classified image sets contain not only comparative image pairs, but also individual images, (3) a concatenation of the high-level global feature of the images and the intermediate local feature of the localized regions is obtained through a pretrained CNN, to support relative attributes prediction on the next stage.

The rest of the paper is organized as follows: some related works are discussed in Sec. 2. In Sec. 3, we describe the proposed method. The experimental setup and results are shown in Sec. 4. Finally, we conclude this paper in Sec. 5.

2 Related Works

2.1 *Binary Attributes*

Attributes based on handcrafted low-level features have shown great success in object classification,[3,4] image search,[7] and object recognition.[10,39] Recent studies show that deep CNN features could achieve a more excellent performance for attribute prediction and attribute-related applications.[19–26] Yang et al.[19] constructed the face descriptors from the different levels of the CNN for different attributes to best facilitate face attribute prediction. Inspired by Yang et al., in this paper, the final deep feature representations for the attributes are formed by the concatenation of the high-level global CNN features and the intermediate local CNN features.

2.2 *Relative Attributes*

Most of the previous works relevant to relative attributes depend on handcrafted features.[7,27,28,40] Recently, deep neural networks have also been extended for ranking applications.[35,36,41] Yaser et al.[35] introduced a CNN-based model, which is composed of a feature learning and extraction part and a ranking part, to predict relative attributes. But it only uses the global deep representations of the images. Krishna and Yong[36] proposed an end-to-end deep neural network to jointly learn the attribute's features, localization, and ranker. They integrate a spatial transformer network (STN) and a ranker network (RN) together in a Siamese network, which is able to localize the relevant image patch corresponding to the visual attribute and train the attribute models simultaneously in a deep learning framework. Though such approach can achieve state-of-the-art performance, it is rather resource demanding. Therefore, our method performs the localization procedure independently in the pipeline.

2.3 *Regions Localization*

Learning attributes based on the relevant attribute regions have shown to produce a superior performance. Most of the existing regions localization approaches rely on pretrained face/body landmark[5] or poselet detectors,[2,37] or crowd-sourcing,[1] and all these methods try to localize binary attributes, whereas our method aims to discover the local regions relevant to relative attributes. The approach of Ref. 30 shares our goal of localizing relative attributes. It uses strongly supervised pretrained facial landmark detectors, and is thus limited to modeling only facial attributes. Moreover, because the detectors are trained independently of the attribute, the learned parts may not necessarily be useful for modeling the desired attribute. Recently, Xiao and Lee[38] proposed a method that discovers the spatial extent of relative attributes automatically by generating and selecting visual chains. This approach directly localizes the attribute without relying on pretrained detectors, and thus can be used to model attributes for any object. However, the images are sorted in the entire image set, the attribute appearance may change not so smoothly among some adjacent images when generating visual chains. Therefore, we propose to first roughly classify the images in the entire image set according to the presence or absence of each attribute, so as to improve the accuracy of the visual chains generation.

3 Proposed Method

3.1 *Regions Localization*

Many previous works have demonstrated that the local feature could achieve a more accurate and informative representation than the global ones. Moreover, many attributes are locally oriented. Therefore, we first localize the image regions that are most relevant to the corresponding attributes.

3.1.1 *Attribute classification*

In this paper, we propose to first coarsely classify the images in the entire image set according to the presence or absence of each attribute. In this way, the accuracy of visual chains generation is improved when discovering the spatial extent of the relative attributes. To this end, we utilize the method of progressive transductive support vector machine (PTSVM) proposed in Ref. 42 to perform the classification task in our work.

For each provided attribute annotation, we first need to label a small set of positive and negative sample images manually. The set of labeled images is denoted as $D_l = \{(x_i, y_i)\}_{i=1}^{l}$, where x_i is the feature vector of image i, $y_i \in \{-1, +1\}$, and the rest of unlabeled images is denoted as $D_u = \{x_i^*\}_{i=l+1}^{n}$. The following minimization problem is optimized over both the separating hyperplane parameters (w, b) and the predicted labels $y^* = (y_{l+1}^*, y_{l+2}^*, \dots, y_n^*)$, $y_i^* \in \{-1, +1\}$

$$\begin{aligned} \min_{w,b,y^*,\xi,\xi^*} & \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{l}\xi_i + C^*\sum_{i=l+1}^{n}\xi_i^*, \\ \text{s.t. } & y_i[w\cdot x_i + b] \geq 1-\xi_i, \quad i = 1,2,\dots,l \\ & y_i^*[w\cdot x_i^* + b] \geq 1-\xi_i^*, \quad i = l+1, l+2,\dots,n \\ & \xi_i \geq 0, \quad i = 1,2,\dots,l \\ & \xi_i^* \geq 0, \quad i = l+1, l+2,\dots,n, \end{aligned} \tag{1}$$

where C and C^* are the user-specified balance parameters. ξ_i and ξ_i^* are the slack variables corresponding to the labeled and unlabeled images, respectively.

When executing the method of PTSVM, all labeled samples are utilized to generate an initial classifier iteratively for each provided attribute annotation. Then, one or two unlabeled samples are labeled using pairwise labeling, i.e., one positive example and one negative example are labeled simultaneously according to Eqs. (2) and (3) for each iteration

$$i_1 = \arg \max_{j:0<f(x_j^*)<1} |f(x_j^*)|, \tag{2}$$

$$i_2 = \arg \max_{j:-1<f(x_j^*)\leq 0} |f(x_j^*)|. \tag{3}$$

The decision function is $f(x) = w\cdot x + b$, and then

$$y_{i_1}^* = \mathrm{sgn}(w\cdot x_{i_1}^* + b), \tag{4}$$

$$y_{i_2}^* = \mathrm{sgn}(w\cdot x_{i_2}^* + b). \tag{5}$$

If there are no samples satisfying one of Eqs. (2) and (3), only one sample is picked and labeled. Meanwhile, all inconsistent labels will be removed by dynamical adjusting.[42] The iterations are performed until all the unlabeled samples are outside the margin band of the separating hyperplane.

Accordingly, we can obtain two image sets for each attribute via attribute classification: S_p for the images with the target attribute, whereas S_q for the images without the target attribute. The entire image set is $S = \{S_p, S_q\}$. Then, we discover the most relevant regions corresponding to an attribute in the two categorized image sets, respectively.

3.1.2 *Regions discovery*

We adapt the method proposed by Xiao and Lee[38] to localize the regions that are most correlated with a target attribute. We modify the ranking method when initializing a visual chain. Given the categorized image sets S_p and S_q corresponding to an attribute, there is a situation as below. For a given comparative image pair (I_i, I_j), I_i is contained in S_p, whereas I_j is contained in S_q, i.e., through classification, the provided comparative image pairs may be separated. That means the categorized image set S_p contains not only the provided image pairs, but also unlabeled separate images. Moreover, we cannot ensure that all the classified image sets contain only the given comparative image pairs, and so far, we have not found a dataset that satisfies this condition. Therefore, we start by sorting the images of S_p in a descending order, using a group sparse-based semisupervised learning approach proposed by Hongxue et al.[28] The ranked image collection is $S_p' = \{I_1', I_2', \dots, I_m'\}$.

To initialize a single chain, we take the top N_{init} images and select one patch from each image to form a patch set $P = \{P_1, P_2, \dots, P_{N_{\text{init}}}\}$. The appearance of each patch varies smoothly with its neighbors in the chain by minimizing the following objective function:

$$\min_P \Phi(P) = \sum_{i=2}^{N_{\text{init}}} \|\phi(P_i) - \phi(P_{i-1})\|_2, \tag{6}$$

where $\phi(P_i)$ is the appearance feature representation of patch P_i in image I_i'. This objective enforces local smoothness. We sample the candidate patches for each image densely at multiple scales. Given the objectives chain structure, we can efficiently find its global optimum using dynamic programming (DP). In the backtracking stage of DP, we can obtain a series of K-best solutions. A chain-level nonmaximum suppression (NMS) is then performed to remove redundant chains and keep a set of K_{init} diverse candidate chains.

After that, we grow each chain along the entire attribute spectrum iteratively by training a detector that adapts to the smoothly changing attribute appearance. To grow the chain, we minimize an objective function again as follows:

$$\min_P \Phi(P) = \sum_{i=2}^{t*N_{\text{iter}}} \|\phi(P_i) - \phi(P_{i-1})\|_2 - \lambda \sum_{i=1}^{t*N_{\text{iter}}} w_t^T \phi(P_i), \tag{7}$$

where λ is a constant that trades off the first local smoothness term and the second detection term. $P = \{P_1, P_2, \dots, P_{t*N_{\text{iter}}}\}$ is the set of patches in a chain. N_{iter} is the number of images considered in each iteration, and w_t is a linear SVM detector learned from the $(t-1)$'th iteration. The same DP is also used here. We repeat the iterative process T times so as to cover the entire attribute spectrum.

As some attribute-relevant regions are hard to detect (e.g., forehead region for "visible forehead"), we can generate new chains by perturbing the existing patches locally in each image with the same perturbation parameters $(\Delta_x, \Delta_y, \Delta_s)$. K_{pert} chains are generated for each of the K_{init} chains with Δ_x and Δ_y each sampled from $[-\delta_{xy}, \delta_{xy}]$ and Δ_s sampled from a discrete set χ, which results in $K_{\text{pert}} \times K_{\text{init}}$ chains in total. The same operations are conducted to the categorized image set S_q, and then the two processed categorized image sets are concatenated together to form the complete chains of the entire image set. There may be an extreme situation, where no comparative image pairs contained in the categorized image set S_q. In such case, we can just make up for some image pairs and remove duplicate images after chains learning randomly. Finally, we rank each chain and select the chains that are mostly correlated with each target attribute.

3.2 *Deep Feature Extraction and Ranking*

After regions localization, we feed both the images and the selected image patches into a pretrained CNN model to obtain the final feature representations. As described by Zhong et al.,[19] the intermediate output of the last convolutional layer could be more effective in specifying shape and variation for the patches that are relevant to an attribute. Therefore, the final deep feature representation is to be the concatenation of the local feature extracted from the last convolutional layer and the global feature output from the last fully connected (FC) layer in this paper. Then, the final deep feature representations are served as the inputs of the ranking module for the task of relative attributes prediction.

In our experiments, we adapt the main deep CNN architecture proposed by Yaser et al.[35] for predicting relative attributes. Similarly, we use a VGG-16[12] model without the last FC layer, which can better satisfy our experimental conditions and experimental requirements. The VGG-16 model contains 13, 3×3 convolutional layers, with max-pooling layers in between and followed by two FC layers. In addition, we apply extra max-pooling steps on the top of convolutional layers to reduce the dimension of intermediate representations (see Fig. 2). Our ranking module is the same as the RankNet proposed in Ref. 35. In the RankNet, the extracted CNN features go through the ranking layer that is a fully connected neural network layer to output the estimated ranks r_i and r_j, for a comparative image pair (I_i, I_j). Then, the estimated ranks r_i and r_j are combined to compute an estimated posterior probability p_{ij}. Finally, the estimated posterior probability p_{ij}, along with the corresponding target probability t_{ij}, is used to calculate the loss, which is then backpropagated to update the weights of the whole network. (See Ref. 35 for more details.)

The illustration of the training process is shown in Fig. 2. Each relative attribute is trained separately. The proposed network takes as input a pair of images (I_i, I_j) and the corresponding local regions that most agree with the relative attribute we are training for. The corresponding target probability t_{ij} according to ground-truth attribute strength is also fed into the ready-made ranking network. Here, t_{ij} is selected from $\{0,0.5,1\}$. If the attribute strength of I_i is greater than that of I_j, then t_{ij} is expected to be >0.5, and vice versa. Furthermore, if the attribute strengths of I_i and I_j are similar to each other, t_{ij} is expected to be 0.5. As shown in Fig. 2, I_i is more smiling than I_j, thus $t_{ij} = 1$. The pair of images and their corresponding patches then go through the deep feature extraction module to obtain the final feature vectors $\phi(I_i)$ and $\phi(I_j)$, respectively. The generated deep representations are later serve as the inputs of the RankNet to compute the loss. Then, the loss is backpropagated to update the weights of each layer.

During the testing (Fig. 3) process, the input is consisted of a single image I_k and the corresponding attribute-correlated local part, whereas the output is the estimated absolute

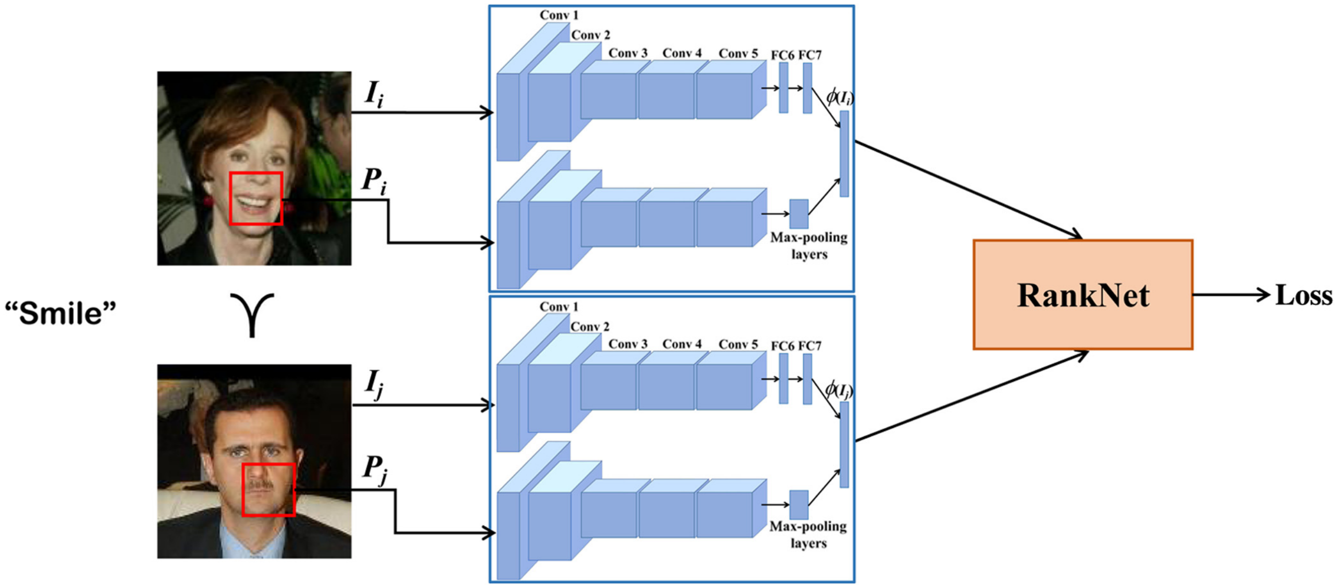

Fig. 2 The schematic view for training. The inputs to our network are a pair of images (I_i, I_j) and their localized regions that most agree with the relative attribute we are training for, as well as corresponding target probability according to the ground-truth attribute strength.

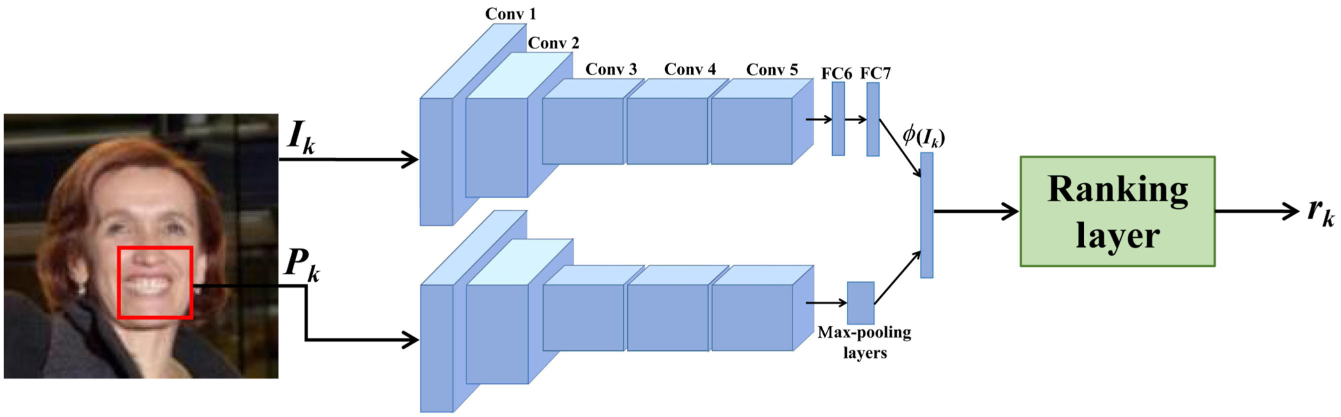

Fig. 3 The schematic view for testing. The input image I_k and the localized most relevant region P_k corresponding to the attribute "smile" go through the deep feature extraction network, and the ranking layer uses the combined features of the local region and image I_k to estimate the absolute rank r_k.

rank r_k for the testing image I_k. According to the estimated absolute ranks, the images set can be ranked easily in the testing image.

4 Experiments

In this section, we quantitatively compare our proposed method with some state-of-the-art methods. Furthermore, we perform multiple qualitative experiments to demonstrate the superiority of our proposed method.

4.1 *Datasets*

Our experiments are evaluated on three public datasets: LFW-10,[30] Zappos50K-1,[29] and Shoes.[32]

LFW-10[30] is a subset of the Labeled faces in the wild (LFW) dataset, which has 2000 images (1000 for training and 1000 for testing) and 10 attribute annotations, with 500 pairs of training and testing images per attribute. The attributes labeled in LFW-10 are "bald head," "dark hair," "eyes open," "good looking," "masculine looking," "mouth open," "smile," "visible teeth," "visible forehead," and "young". In our experiments, we follow the training/testing split of Ref. 30.

Zappos50K-1 is a subset of the UT-Zap50K dataset,[29] which provides 1388 training and 300 testing pairs on average for each of the four attributes: "open," "sporty," "pointy," and "comfort." We use the same training/testing split as that in Ref. 29. Shoes[32] dataset contains 14658 shoe images and 10 attributes, of which three are overlapping with Zappos50K-1: "open," "sporty," and "pointy." Because there are only about 140 pairs of relative attribute annotations per attribute, we use this dataset only for testing.

4.2 *Experimental Setup*

The evaluation is performed on a platform with GTX 1060 GPU (6G memory), 3.3 GHz CPU, and 32 GB memory. The image features utilized for attribute classification and initial ranking of the categorized image sets are represented by a concatenation of GIST descriptors and LAB color histograms.[27,28] For attribute classification, we label five positive examples and five negative examples for each attribute in both LFW-10 and Zappos50K-1 training sets, and implement PTSVM based on Joachims's SVM$^{\text{light}}$.[43] The constants C and C^* are set to 1 and 0.5, respectively. To sort the images in the categorized image sets, the setting is similar to that in Ref. 28 except modifying the parameter d to 9.2. Furthermore, we use HOG features for detection and local smoothness, and set $N_{\text{init}} = 5$, $N_{\text{iter}} = 60$, $\lambda = 0.05$, $K_{\text{init}} = 20$, $K_{\text{pert}} = 15$, $\delta_{xy} = 0.6$, $\chi = \{1/4, 1\}$, and $T = 3$.

For the deep feature extraction part, we initialize the weights using the pretrained model on ILSVRC 2014[44] for the task of image classification. Extra max-pooling layers are appended to the fifth pooling layer to reduce the dimension of intermediate representations. For the ready-made ranking part, we initialize the weights w of the ranking layer using the Xavier method,[45] and initialize the bias to 0. During training, we use a mini-batch size of 16 image pairs for SGD, and train is done after 50 and 30 epochs for LFW-10 and Zappos50K-1 datasets, respectively. The initial learning rates of the deep feature extraction layers and the ranking layer are set to 10^{-5} and 10^{-4}, respectively, and then are dynamically changed by RmsProp.[46] Moreover, the estimated posterior p_{ij} of the ranking network is restricted in $[10^{-6}$, 1 to $10^{-6}]$ to prevent the binary cross entropy loss from diverging.

4.3 *Quantitative Results*

In this paper, we report the accuracy in terms of the percentage of correctly ordered image pairs, and the comparative data are collected from previous works.

Figure 4 shows the results on LFW-10 dataset. We can see that our method using only the high-level local CNN feature performs better on the locally orientated attributes, such as "mouth open," "smile," which demonstrates that our regions localization module is more efficient than that of Ref. 38. Moreover, as shown in Fig. 4, we produce the best results on six of the 10 attributes.

Figure 5 shows the results on Zappos50K-1 dataset. Our method achieves the state-of-the-art accuracy again. As the shoe images in this dataset are well aligned, centered, and have clear backgrounds, we can obtain a high accuracy. It is observed that the improvement over the abstract attribute "comfort" is slight, whereas the improvements are more remarkable over the locally orientated attributes, such as "open" and "pointy." The ranking accuracy comparison of Ref. 38 with both global and local CNN features and our method all with CNN feature demonstrates that deep feature

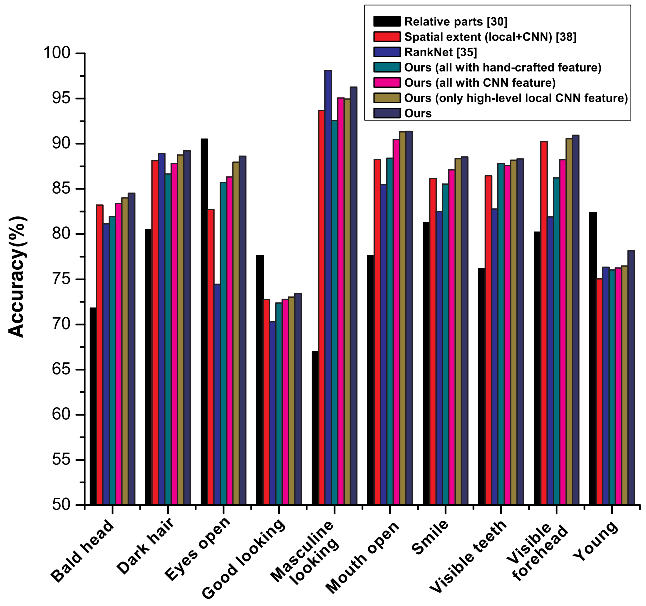

Fig. 4 Comparison of ranking accuracy on LFW-10 dataset.

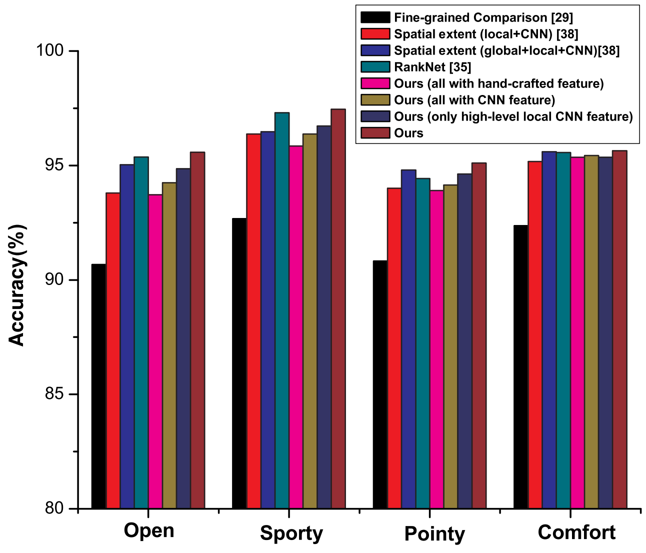

Fig. 5 Comparison of ranking accuracy on Zappos50K-1 dataset.

did not contribute enough to regions localization in this paper.

Figure 6 shows our results on the Shoes dataset. We take our models trained on Zappos50K-1, and test on Shoes to evaluate cross-dataset generalization ability. Figure 6 shows the comparison results of the three overlapping attributes ("open," "pointy," and "sporty") contained in both Zappos50K-1 and Shoes datasets, respectively. Compared with other methods using CNN feature, our method all with handcrafted feature obviously performs the worst.

Table 1 shows the mean ranking accuracy of the corresponding methods on LFW-10 (see Fig 4), Zappos50K-1 (see Fig 5), and Shoes dataset (see Fig 6). On the LFW-10 dataset, our mean accuracy is 2.27% and 4.75% higher than that of Refs. 38 and 35, respectively. Although all the corresponding methods can achieve a high mean ranking accuracy on the Zappos50K-1 dataset, our approach performs the best. For the three overlapping attributes of the Shoes dataset, we just obtain a slight improvement of 0.09% absolute over the method of Singh and Lee.[36]

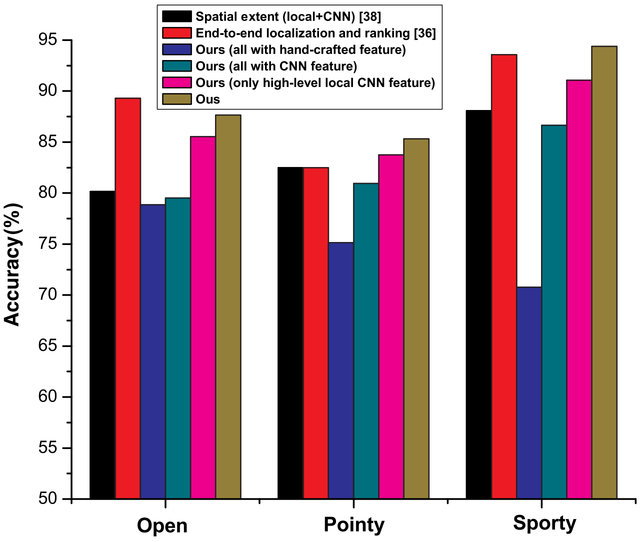

Fig. 6 Comparison of ranking accuracy on Shoes dataset using the models trained on Zappos50K-1 dataset. The result demonstrates the cross-dataset generalization ability of our method.

Table 1 Mean ranking accuracy of the corresponding methods on LFW-10, Zappos50K-1, and Shoes dataset.

	LFW-10	Zappos50K-1	Shoes
Relative parts[30]	78.50		
Fine-grained comparison[29]		91.64	
Spatial extent (local + CNN)[38]	84.66	94.83	83.58
Spatial extent (global + local + CNN)[38]		95.47	
RankNet[35]	82.18	95.67	
End-to-end localization and ranking[36]			88.46
Ours (all with handcrafted feature)	84.32	94.71	74.91
Ours (all with CNN feature)	85.50	95.05	82.37
Ours (only high-level local CNN feature)	86.36	95.39	86.78
Ours	**86.93**	**95.88**	**88.55**

4.4 *Qualitative Results*

Figure 7 shows the sample results of the global ranking on the LFW-10 test images. Each row corresponds to a face attribute and exhibits decreasing attribute strength. It can be observed that, for the locally orientated attributes such as "mouth open," "smile," the results are basically visually correct. Although for the more global attributes, such as

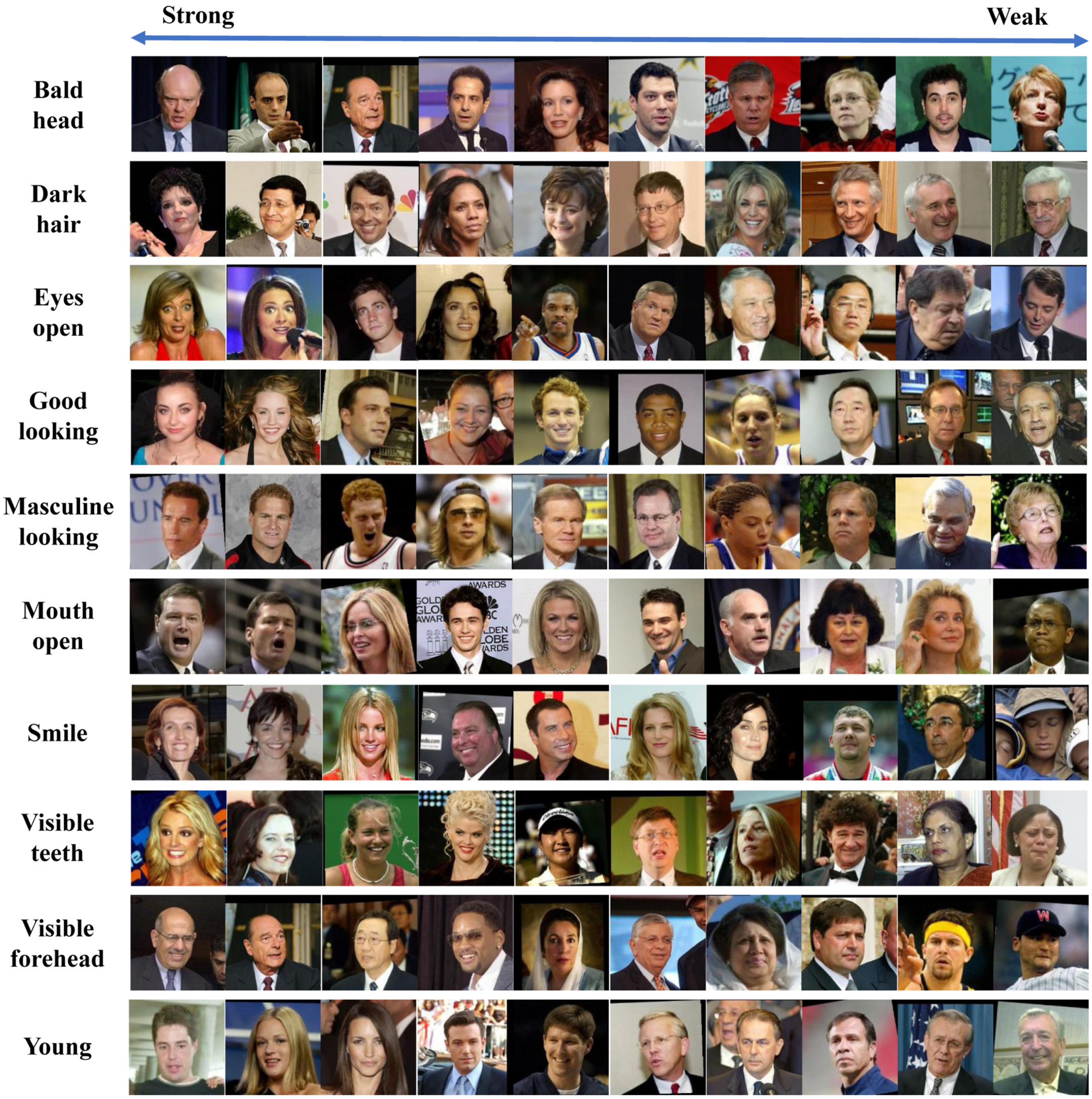

Fig. 7 Sample results of the global ranking on the LFW-10 test images. Each row corresponds to a face attribute and exhibits decreasing attribute strength. It is shown that the ranking obtained by our method is accurate for all attributes.

"masculine looking," there are more visual mistakes. Thus, it can be seen that the locally orientated attributes benefit more from our work.

Figure 8 shows the sample ranking results for the four provided attributes on the Zappos50K-1 test images. The results demonstrate that our method is capable of generating accurate image rankings using the attribute-correlated local patches and their corresponding intermediate CNN features.

4.5 *Ablation Study*

We study the contribution of the two operations that use either only the attribute classification step or only the intermediate local CNN features extraction step on the ranking performance. When conducting only the attribute classification step, the final deep representations are the combination of both the global and local CNN features from the last FC layer.

Table 2 shows the attribute ranking accuracy of the two separate operations, as well as that of our combined method on LFW-10. It can be observed that the attribute classification baseline contributes more than the intermediate local output baseline to the ranking performance. The intermediate local output baseline may weaken the accuracy improvements of attributes that are global, such as "masculine

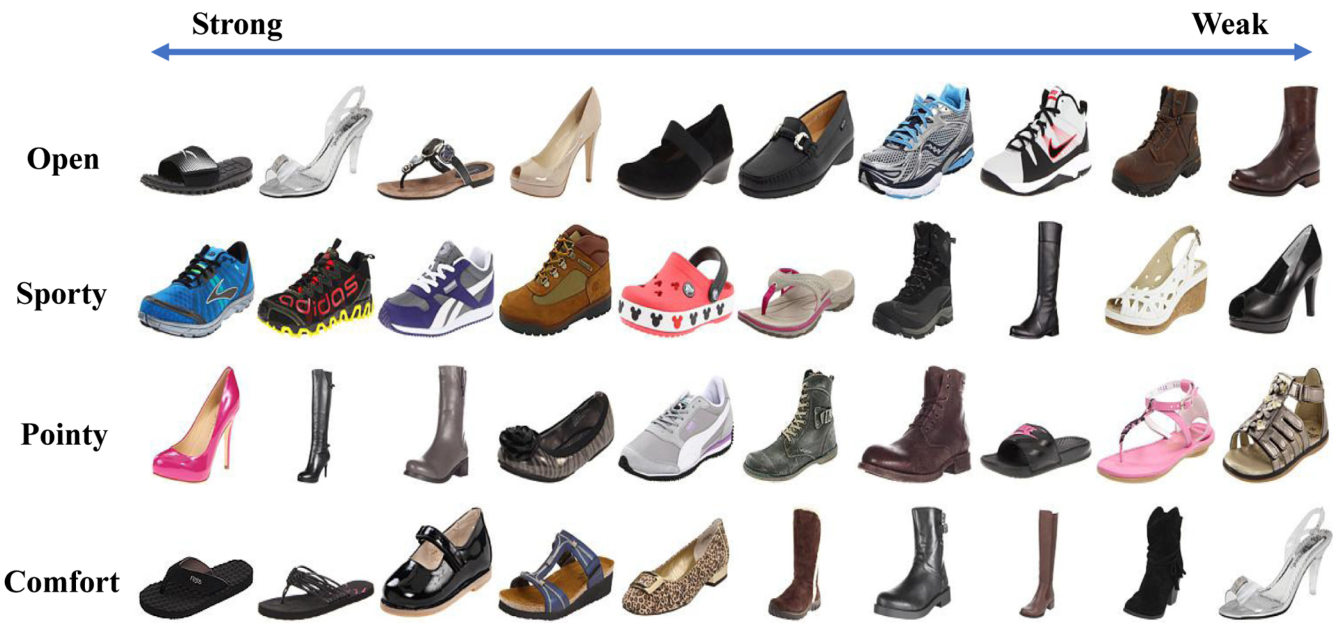

Fig. 8 Sample ranking results for the four provided attributes on the Zappos50K-1 test images. The ranking is also accurate for each attribute.

Table 2 Attribute ranking accuracy of the two separate operations, as well as that of our combination method on LFW-10 dataset.

	B head	D hair	Eyes O	G looking	M looking	Mouth O	Smile	V teeth	V forehead	Young	Mean
Attribute classification	84.33	**89.27**	88.45	73.26	**96.34**	91.25	88.29	88.04	90.69	**78.28**	86.82
Intermediate local output	83.69	88.38	87.71	72.82	95.26	90.59	87.94	87.62	90.43	76.25	86.07
Combined	**84.52**	89.20	**88.62**	**73.43**	96.25	**91.38**	**88.52**	**88.31**	**90.93**	78.16	**86.93**

looking" and "young." The third row in Table 2 shows the result of our combined method, which produces the best accuracy for seven out of the 10 attributes.

4.6 *Application to Interactive Image Search*

Relative attributes not only help to describe a pair of images more clearly but also help to retrieve images more carefully. Similar to the feedback collection setup of Ref. 30, we perform the interactive image search using relative attribute-based feedback, which is a significant application of relative attributes. Given a target image, it can be described through attribute's feedbacks with respect to a few reference images. The search set is divided into two disjoint sets according to a given feedback with respect to a reference image. The rank of all the images in the search set is averaged over all feedbacks with respect to all reference images, using absolute classifier score difference. We calculate the number of the predicted target images falling below a given rank, and more search images mean better performance. We use the LFW-10 testing dataset as our search set. The number of relative attribute-based feedbacks is varied in {2,5,10} corresponding to one or two reference images. Table 3 shows the number of search images corresponding to different settings, based on a total of 275 searches per setting. The first column shows the specified image rank. It can be observed that the number of search images raises with an increase in the number of feedbacks and/or number of reference images.

Table 3 The number of search images corresponding to different settings on LFW-10 testing dataset.

	One reference image			Two reference images		
	Two feedbacks	Five feedbacks	Ten feedbacks	Two feedbacks	Five feedbacks	Ten feedbacks
100	53	83	94	57	88	98
200	86	118	141	97	136	165
300	120	159	173	128	179	201

Our result outperforms that of Ref. 30 by 18 search images on average.

5 Conclusion

In this paper, we propose the deep relative attributes learning strategy, which is implemented based on conventionally acquired attribute-correlated local regions. We first perform attribute classification rather than discovering the spatial extents corresponding to each provided attribute over the entire image set directly. In this way, the images and localized regions are both fed into the pretrained CNN model. The final outputs are the concatenation of last global features and intermediate local features, which were used to predict relative attributes. On three public relative attribute prediction benchmarks, we show that the proposed attribute classification procedure is an effectiveness way for learning attribute relevant local regions. However, for side face images, we still could not learn the local regions to certain attributes effectively. We want to impose some constraints on the learned local regions, which is the problem we need to solve in the future work.

Disclosures

The authors declare no conflict of interest.

Acknowledgments

This research was supported by the Foundation for Innovative Research Groups of the NSFC (Grant No. 71421001), and the National Natural Science Foundation of China (Grant Nos. 61772111 and 61502073).

References

1. K. Duan et al., "Discovering localized attributes for fine-grained recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (2012).
2. N. Zhang et al., "Panda: pose aligned networks for deep attribute modeling," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1637–1644 (2014).
3. Z. Akataa et al., "Label-embedding for attribute-based classification," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 819–826 (2013).
4. C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Trans. Pattern Anal. Mach. Intell.* **36**, 453–465 (2014).
5. N. Kumar et al., "Attribute and simile classifiers for face verification," in *IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 365–372 (2009).
6. N. Kumar et al., "Describable visual attributes for face verification and image search," *IEEE Trans. Pattern Anal. Mach. Intell.* **33**(10), 1962–1977 (2011).
7. A. Kovashka and K. Grauman, "Attribute adaptation for personalized image search," in *IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 3432–3439 (2013).
8. L. An et al., "Scalable attribute-driven face image retrieval," *Neurocomputing* **172**, 215–224 (2016).
9. R. Deshmukh Hema et al., "Scalable face image retrieval using attribute patch reinforcement and sparse coding," *Int. J. Eng. Sci.* **6**(3), 2697–2700 (2016).
10. A. Farhadi et al., "Describing objects by their attribute," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1778–1785 (2009).
11. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Twenty-Sixth Annual Conf. on Neural Information Processing Systems (NIPS)*, pp. 1097–1105 (2012).
12. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv: 1409.1556 (2014).
13. C. Szegedy et al., "Going deeper with convolutions," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9 (2015).
14. K. He et al., "Deep residual learning for image recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778 (2016).
15. D. Zhu et al., "Image salient object detection with refined deep features via convolution neural network," *J. Electron. Imaging* **26**(6), 063018 (2017).
16. X. Liu et al., "Adaptive metric learning with deep neural networks for video-based facial expression recognition," *J. Electron. Imaging* **27**(1), 013022 (2018).
17. P. Li et al., "Deep convolutional computation model for feature learning on big data in internet of things," *IEEE Trans. Ind. Inf.* **14**(2), 790–798 (2018).
18. Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in *IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 2590–2600 (2017).
19. Y. Zhong, J. Sullivan, and H. Li, "Face attribute prediction using off-the-shelf CNN features," arXiv: 1602.03935 (2016).
20. C. Huang, C. Change Loy, and X. Tang, "Unsupervised learning of discriminative attributes and visual representations," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 5175–5184 (2016).
21. Z. Liu et al., "Deep learning face attributes in the wild," in *IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 3730–3738 (2015).
22. V. Escorcia, J. C. Niebles, and B. Ghanem, "On the relationship between visual attributes and convolutional networks," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1256–1264 (2015).
23. A. S. Razavian et al., "CNN features off-the-shelf: an astounding baseline for recognition," arXiv: 1403.6382 (2014).
24. Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2892–2900 (2015).
25. J. Shao et al., "Deeply learned attributes for crowded scene understanding," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 4657–4666 (2015).
26. Z. Liu et al., "Deepfashion: powering robust clothes recognition and retrieval with rich annotations," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1096–1104 (2016).
27. D. Parikh and K. Grauman, "Relative attributes," in *IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 503–510 (2011).
28. H. Yang et al., "Semi-supervised learning based on group sparse for relative attributes," in *IEEE Int. Conf. on Image Processing (ICIP)*, pp. 3931–3935 (2015).
29. A. Yu and K. Grauman, "Fine-grained visual comparisons with local learning," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 192–199 (2014).
30. R. N. Sandeep, Y. Verma, and C. V. Jawahar, "Relative parts: distinctive parts for learning relative attributes," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3614–3621 (2014).
31. L. Liang and K. Grauman, "Beyond comparing image pairs: setwise active learning for relative attributes," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 208–215 (2014).
32. A. Kovashka, D. Parikh, and K. Grauman, "Whittlesearch: image search with relative attribute feedback," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2973–2980 (2012).
33. A. Kovashka and K. Grauman, "Attribute pivots for guiding relevance feedback in image search," in *IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 297–304 (2013).
34. L. Chen, Q. Zhang, and B. Li, "Predicting multiple attributes via relative multi-task learning," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1027–1034 (2014).
35. Y. Souri, E. Noury, and E. Adeli, "Deep relative attributes," in *Asian Conf. on Computer Vision (ACCV)*, pp. 118–133 (2016).
36. K. K. Singh and Y. J. Lee, "End-to-end localization and ranking for relative attributes," in *European Conf. on Computer Vision (ECCV)*, pp. 753–769 (2016).
37. L. Bourdev, S. Maji, and J. Malik, "Describing people: a poselet-based approach to attribute classification," in *IEEE Int. Conf. on Computer Vision (ICCV)* (2011).
38. F. Xiao and Y. J. Lee, "Discovering the spatial extent of relative attributes," in *IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 1458–1466 (2015).
39. R. Tao, A. W. Smeulders, and S.-F. Chang, "Attributes and categories for generic instance search from one example," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 177–186 (2015).
40. S. Li, S. Shan, and X. Chen, "Relative forest for attribute prediction," in *Asian Conf. on Computer Vision (ACCV)*, pp. 316–327 (2012).
41. Y. Song, H. Wang, and X. He, "Adapting deep ranknet for personalized search," in *ACM Int. Conf. on Web Search and Data Mining*, pp. 83–92 (2014).
42. Y. Chen, G. Wang, and S. Dong, "Learning with progressive transductive support vector machine," *Pattern Recognit. Lett.* **24**, 1845–1855 (2003).
43. T. Joachims, "SVM-light support vector machine," http://svmlight.joachims.org/ (2008).
44. O. Russakovsky et al., "Imagenet large scale visual recognition challenge," *Int. J. Computer Vision* **115**(3), 211–252 (2015).

45. X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Thirteenth Int. Conf. on Artificial Intelligence and Statistics, PMLR*, Vol. 9, pp. 249–256 (2010).
46. T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude," in *COURSERA: Neural Networks for Machine Learning* (2012).

Fen Zhang is a PhD student at Dalian University of Technology. She has earned a bachelor's degree in Electronic Science and Technology in 2005, and her master's degree in signal and information processing (SIP) from JiangSu University of Science and Technology in 2008 respectively. After that, she engaged in work related to software programming. She began to pursue PhD degree from September 2012 and her main research trend at present is attribute-based image retrieval.

Xiangwei Kong is a professor at Dalian University of Technology. She received her PhD from Dalian University of Technology in 2003. She was a visiting research scholar at Purdue University from September 2006 to September 2007 and at NYU from December 2014 to June 2015. Her research interests include multimedia forensics, pattern recognition, and information retrieval.

Ze Jia is a naval officer and his main research trend at present is signal and information processing.