FASHION ANALYSIS WITH A SUBORDINATE ATTRIBUTE CLASSIFICATION NETWORK

Huijing Zhan[1], *Boxin Shi*[2], *Alex C. Kot*[1]

[1]School of Electrical and Electronic Engineering, Nanyang Technological University
[2]Artificial Intelligence Research Center, National Institute of AIST
{hjzhan,eackot}@ntu.edu.sg, boxin.shi@aist.go.jp

ABSTRACT

In this paper we deal with two image-based object search tasks in the fashion domain, clothing attribute prediction and cross-domain shoe retrieval. Clothing attribute prediction is about describing the appearances of clothes via semantic attributes and cross-domain shoe retrieval aims at retrieving the same shoe items from online stores given a daily life shoe photo. We jointly solve these two problems by a novel Subordinate Attribute Convolutional Neural Network (SA-CNN), with the newly designed loss function that systematically merges semantic attributes of closer visual appearance to prevent images with obvious visual differences being confused with each other. A three-level feature representation is further developed based on SA-CNN for shoes from different domains. The experimental results demonstrate that the clothing attribute prediction using the proposed SA-CNN achieves better performance than that using traditional features and fine-tuned conventional CNN. Moreover, for the task of cross-domain shoe retrieval, the top-20 retrieval accuracy with deep features extracted from SA-CNN has a significant improvement of 43% compared to that with the pre-trained CNN features.

Index Terms— Fashion analysis, subordinate attribute, clothing attribute prediction, cross-domain, shoe retrieval,

1. INTRODUCTION

In recent years, the explosive growth of the online shopping brings considerable sales in the fashion domain, among which clothing and shoes make up the very large proportion. Driven by their huge profits, visual analysis of fashion is recently receiving extensive attention in the multimedia community. For example, the vision-based technique that automatically predicts clothing attributes can save much labor for online clothing sellers. It is also desired in the crime investigation scenario, which is capable of generating clothing descriptions and searching the criminals by matching the descriptions provided by the witness. Another situation often occurring in our shopping experience is that we may want to look for exactly the same shoe item from online stores when seeing a pair of beautiful shoes on the shop window. However, the text-based search engine usually fails to provide satisfying results due to the limited descriptive capability of several words. Hence it is necessary to develop a visual retrieval system in such scenarios.

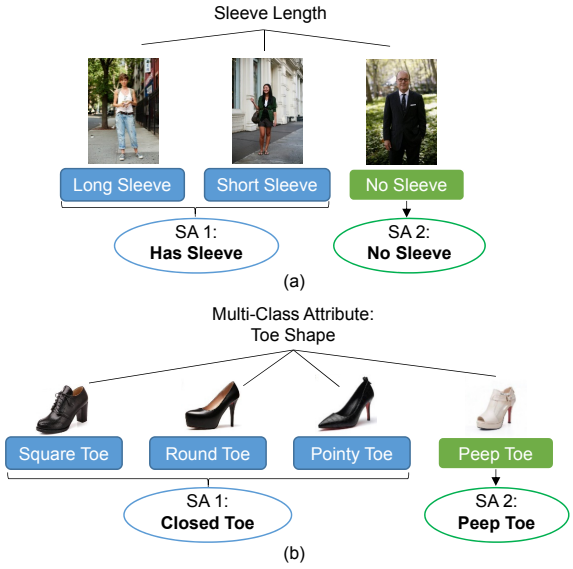

Fig. 1. The attribute appearances of the multi-class attribute "Sleeve Length" are illustrated in (a). It is further merged into two subordinate attribute classes. The same applies to "Toe Shape" attribute as shown in (b), merged into "Closed Toe" (SA1) and "Peep Toe" (SA2) denoted in blue ellipse.

The above mentioned two problems draw great interest among multimedia researchers these days. The first problem is called *clothing attribute prediction* and the second problem is referred as *cross-domain shoe retrieval*. The shoe images captured in the daily life environment is named as *street domain* while online shop pictures as *online domain*.

*This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by the National Research Foundation, Singapore, under its Interactive Digital Media (IDM) Strategic Research Programme. Boxin Shi is supported by a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

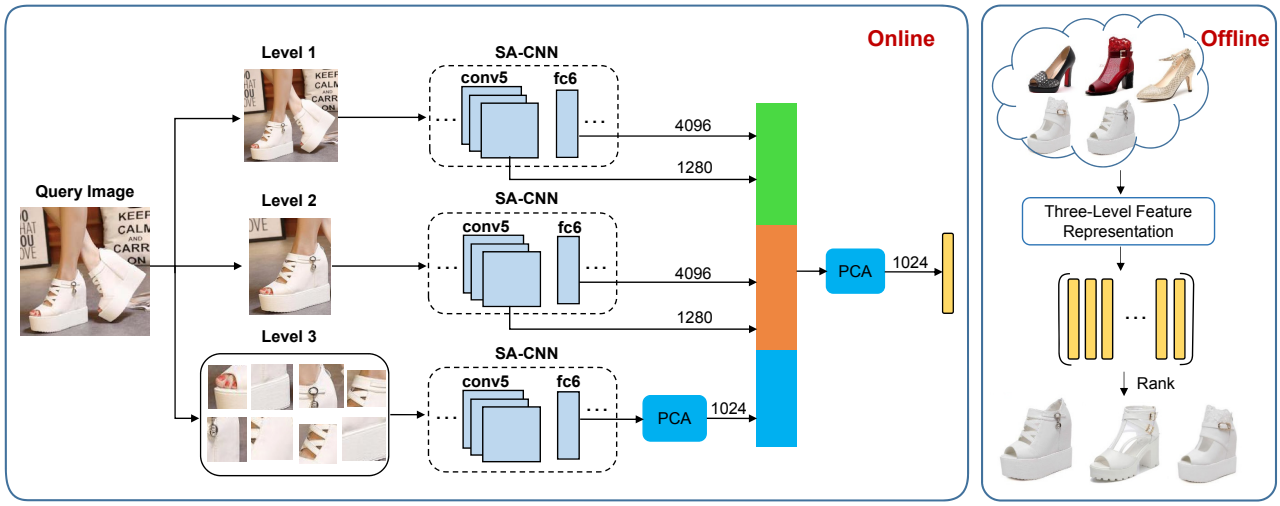

Fig. 2. The pipeline of our proposed shoe retrieval system.

Existing works on attribute prediction of clothes (*e.g.,* [1]) investigated the mutual dependencies between the attributes with a Conditional Random Field [2]. However, to the best of our knowledge, the inherent properties of the multi-class attribute have not been fully explored. We further organize semantic attributes in a hierarchical manner as Subordinate Attributes (SA), defined as merging semantic attributes with closer visual appearances. The correct prediction of subordinate attribute is always given higher priority to ensure that clothes or shoes with obviously different attributes are never confused with each other. Thus, a new loss function is defined to penalize the misclassification between different subordinate attributes by adding an extra regularization term in addition to the typical softmax-loss term. The concept of the subordinate attribute for "Sleeve Length" and "Toe Shape" is illustrated in Fig. 1. For the semantic attribute prediction of clothes, the proposed SA-CNN is directly applied to the entire image and the performance is evaluated on the clothing attributes dataset released in [1].

For the cross-domain shoe retrieval, an effective descriptor is of vital importance, which is expected to not only capture the low-level similarities but also convey the semantic relationships. Unfortunately, due to the large discrepancies of cross-domain shoe images such as self-occlusion, cluttered background, scale variation and viewpoint variation, the commonly adopted descriptors [3, 4, 5] in the retrieval system lose their effectiveness. The most recent work by Kiapour *et al.* [6] presented a three-layer category-independent metric network for the street-to-shop product similarity evaluation. Nevertheless, the features from the fc6 layer of the pre-trained AlexNet [7] on the whole image is not sufficient to depict the appearance of shoes in various scales. This motivates us to develop discriminative shoe feature representations to maximize the distances for different shoes and minimize that for the same shoes in different domains.

Specifically, we incorporate SA-CNN with three levels of input as the deep representation for shoes as illustrated in Fig. 2. The multi-level inputs capture the image content in a coarse-to-fine manner. The first level is the whole image, which conveys the coarsest scene-level semantics of the image. Level 2 and Level 3 contain increasingly finer-level details with the top-1 scored region proposal using our proposed region proposal quality ranking algorithm and semantic shoe part patches detected by Deformable Part Model (DPM) [8]. We utilize the cosine distance to measure the similarity between the query and reference gallery images.

This paper extends our earlier work in shoe retrieval [9]. The hierarchical properties of multi-class attribute is considered in the design of SA-CNN and a novel loss function is developed to penalize the prediction error on subordinate attributes. Moreover, the proposed SA-CNN is applied to the clothing attribute prediction, which verifies its availability on other fashion products in additional to shoes.

2. SA-CNN: SUBORDINATE ATTRIBUTE BASED CONVOLUTIONAL NEURAL NETWORK

As illustrated in Fig. 1, treating different semantic attributes equally loses the discriminative visual features for both clothes and shoes. We propose a hierarchical structure of semantic attribute grouping and a corresponding CNN with a novel loss function to address this issue.

2.1. Network Architecture

Fig. 3 shows the architecture of the proposed SA-CNN. We keep the first five convolutional layers followed by the first fully-connected layer fc6 of the AlexNet [7] and develop cus-

tomized fully-connected fc7 and fc8 layers for each attribute to obtain attribute-aware features. Thus, all the attributes share the weights of the shallow layers and differ in the high-level semantics. For each attribute, its specific fc8 branches out several units, based on the number of potential sub-class attribute values.

2.2. Loss Function

Our attributes mostly belong to the multi-class attribute type, so we use the multi-class softmax loss function. Here we take one attribute a_i as an example and the rest attributes follow the same procedure. Assume the training set contains N images $x_i \in \{1, 2, ..., N\}$ with $y_i \in \{1, 2, ..., K\}$ as the corresponding attribute labels, where K is the number of values for the specified attribute a_i. In the standard CNN, the loss function is minimized in order to maximize the posterior probability for the ground truth class label of the given training samples. Let $h_j^{(i)}(j = \{1, 2, ..., K\})$ indicate the activation value of the node j from the last fully-connected layer (fc8) in Fig. 3, then the probability that the given training sample x_i belongs to class j, denoted as $P_j^{(i)}$, is calculated as:

$$P_j^{(i)} = \frac{\exp(h_j^{(i)})}{\sum_{g=1}^{K} \exp(h_g^{(i)})}. \tag{1}$$

Given the probability of the softmax loss layer $P_j^{(i)}$, the cost function to be minimized becomes

$$J_0 = -\frac{1}{N}\left[\sum_{i=1}^{N}\sum_{j=1}^{K} \delta(y^{(i)} = j) \log P_j^{(i)}\right], \tag{2}$$

where δ is an indicator function. The basic idea behind the loss function is that the prediction error for each class is treated equally. However, in our problem, the classification error made within each subordinate attribute cluster is smaller than the errors between clusters. Thus, the penalty between subordinate classes should be larger than the misclassification made within subordinate attribute class. As mentioned above, the sub-class attribute values can be further categorized into subordinate attribute classes by merging the semantically-close attribute values into large clusters. Assuming that we have the K attribute values which can be further categorized into M subordinate attribute classes, then we use $\{\mathbb{G}_1, \mathbb{G}_2, ..., \mathbb{G}_M\}$ to represent M semantically-close clusters, where M is the number of subordinate attribute classes (clusters). Here $\mathbb{G}_t(t = 1, 2, ..., M)$ is a real-valued set containing the attribute values in the t-th cluster. The proposed loss function based on the property of subordinate attribute is formulated as:

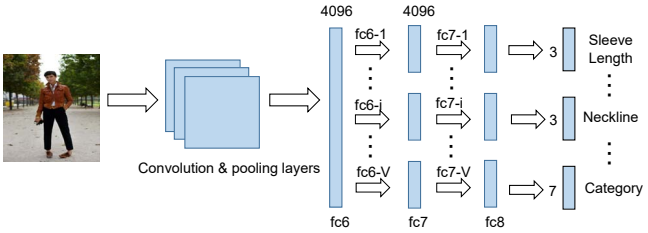

Fig. 3. The architecture of the proposed SA-CNN. Here we assume we have V types of semantic attributes.

$$J = J_0 - \frac{\lambda}{N}\Bigg[\sum_{i=1}^{N}\Bigg(\sum_{j\in\mathbb{G}_1} \delta(y^{(i)} = j) \log P_{\mathbb{G}_1}^{(i)} + \sum_{j\in\mathbb{G}_2} \delta(y^{(i)} = j) \log P_{\mathbb{G}_2}^{(i)} + ... + \sum_{j\in\mathbb{G}_M} \delta(y^{(i)} = j) \log P_{\mathbb{G}_M}^{(i)}\Bigg)\Bigg]. \tag{3}$$

It can be seen that the probability for the same cluster is shared. The probability of a given sample x_i belonging to the cluster $\mathbb{G}_t$ is denoted as below:

$$P_{\mathbb{G}_t} = \sum_{j\in\mathbb{G}_t} P_j^{(i)}, \tag{4}$$

where λ is a regularization parameter controlling the importance between the original loss term J_0 and the subordinate attribute classification term. As we can see, if we set λ as a small value, then the subordinate classification loss term makes slight contribution; otherwise, the effect of sub-class partition is very weak. We set λ as 2 in our experiments based on the validation data.

The partial derivatives of Eq. 3 with respect to the output of the fc8 layer $(h_j^{(i)}; j = \{1, 2, ..., K\})$ is computed and the Stochastic Gradient Descent (SGD) algorithm is utilized to update the parameters for the network. The partial derivatives of J_0 is provided in the literature [10], which are shown as follows:

$$\frac{\partial J_0}{\partial h_j^{(i)}} = \frac{1}{N}\left[P_j^{(i)} - \delta(y^{(i)} = j)\right]. \tag{5}$$

Different from the partial derivative of the original softmax loss function, the partial derivative of the term J with respect to the activation value of fc8 layer is based on the different input. Assume that $y^{(i)} \in \mathbb{G}_t$, then the partial derivatives of the new loss J are calculated as:

$$\frac{\partial J}{\partial h_{j\notin\mathbb{G}_t}^{(i)}} = \frac{1}{N}\left[(\lambda + 2)P_j^{(i)} - 2\delta(y^{(i)} = j)\right], \tag{6}$$

and

$$\frac{\partial J}{\partial h^{(i)}_{j\in\mathbb{G}_t}} = \frac{1}{N}\left[(\lambda + 2 - \frac{\lambda}{P_{\mathbb{G}_t}})P^{(i)}_j - 2\delta(y^{(i)} = j)\right]. \tag{7}$$

3. FEATURE REPRESENTATION FOR SHOES

Besides achieving a better clothing attribute prediction performance based on the hierarchical groupings of the multi-class attributes, the proposed SA-CNN can also be applied to obtain a more semantic feature representation for shoes to facilitate the task of cross-domain shoe retrieval. In this section, we first introduce the novel region proposal quality ranking algorithm followed by the multi-level feature representation for shoes.

We consider the localization of the shoes equivalent to a region proposal quality ranking procedure. Existing popular proposal generation methods like EdgeBox [11] and Selective Search [12] produce an initial set of region proposals based on different low-level cues, *e.g.*, the similarity of the neighboring super-pixels or the number of contours included, which contains many noisy candidate windows with low Intersection-Over-Union (IoU) scores. For the generated region proposal r_i and the ground truth annotations g_i, IoU is defined as:

$$\text{IoU}(r_i, g_i) = \frac{\gamma(r_i \cap g_i)}{\gamma(r_i \cup g_i)}, \tag{8}$$

where $\gamma(\cdot)$ computes the area of the specified region. Thus, it is necessary to re-rank the initial pool of region proposals and the proposal with the highest rank is returned to be the true location of shoes.

We leverage the advantages of different top-performing models and develop our region proposal quality measurement approach. In our work, the quality of the region proposal is evaluated in terms of three mid-level cues: the probability score from CNN detection model (c), the confidence score by DPM (d) and the objectiveness score returned by EdgeBox (e). Then rankSVM [13] is utilized to learn the weights balancing the importance of three scores. For the CNN detection model, it computes a probability vector and assigns a binary foreground/background label to each region proposal.

With the ordered pairs $\mathbb{O}$ and their pairwise labels y, each region proposal is represented by $\mathbf{h}$. The goal is to learn a mapping function $f(\mathbf{h}) = \mathbf{w}^\top \mathbf{h}$, which predicts its corresponding quality score and estimates the relationship between data pairs (s_k, t_k) with the following constraint:

$$\forall u_{s_k} > u_{t_k} : f(\mathbf{h}_{s_k}) > f(\mathbf{h}_{t_k}). \tag{9}$$

The rankSVM model is leveraged by minimizing the objective function:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{(s_k,t_k)\in\mathbb{O}} L(\mathbf{w}^T\mathbf{h}_{s_k} - \mathbf{w}^T\mathbf{h}_{t_k}), \tag{10}$$

where C is the trade-off parameter and L is a loss function with the form $L(t) = \max(0, 1-t)$.

A query image q from the street domain is represented as $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_P]$ with the confidence scores computed from the P initially produced region proposals. Then the quality scores are indicated as $\mathbf{J} = \mathbf{w}^\top \mathbf{H}$, where $\mathbf{J}$ is P-dimensional vector and each element indicates the quality score for the corresponding region proposal. The top-1 scored region proposal is selected as the Level 2 image and the activated feature from the conv5 and fc6 layers of the SA-CNN is represented as $F_{L2} \in \mathbb{R}^{5376}$. Here we employ the two level pyramid mean-pooling [14] to the conv5 feature map. Also, the Level 1 feature $F_{L1} \in \mathbb{R}^{5376}$ is extracted in the similar manner. The Level 3 feature are represented by the stacked fc6 activated features using the fine-grained part patches followed by the PCA operation, denoted as $F_{L3} \in \mathbb{R}^{1024}$ (See Fig. 2).

4. EXPERIMENTS AND RESULTS

4.1. Datasets

Clothing Attributes Dataset: It is composed of 1856 images with 3 multi-class attributes and 23 annotated binary attributes. More details of the dataset are provided in [1]. In this work we are only interested in intrinsic properties of the multi-class attributes and ignore the binary attributes.

Shoe Dataset: To evaluate the performance of our cross-domain shoe retrieval approach, we collect a novel shoe dataset. It consists of about 5820 shoe pictures from the online domain and 8020 daily shoe photos from the street domain. A unique product ID is assigned to each shoe image indicating its shoe model index with 11 types of fine-grained semantic shoe attributes. Each online shoe photo contains an individual shoe facing to the left side 45 degree view and clean background. We crawl these images from *Jingdong*[1]. Each of them has several daily life photos and their counterparts from the online domain. To our best knowledge, we have established the first cross-domain shoe image dataset with semantic attribute annotations.

4.2. Experimental Settings

For the experiments of predicting clothing attributes, we follow the same training and testing split as reported in [1]. Simple augmentation techniques such as cropping, flipping, *etc* are applied to generate more images for SA-CNN training.

With respect to the training and testing data used in the cross-domain shoe retrieval system, different SA-CNN models for the street and online domains are learnt separately. About 15000 scored region proposals are obtained using 3000 street shoe photos with the top-5 scored proposal candidates and they are fed into the street domain SA-CNN model for

[1]https://www.jd.com/

Fig. 4. (a) Example street and online domain images from the Shoe Dataset; (b) Photos in the Clothing Attributes Dataset together with the semantic attribute annotations.

training. Meanwhile, the online domain SA-CNN is learnt using the 18000 online shoe photos (after data augmentation).

Next we introduce the models involved in the Level 2 feature generation process. About 2600 shoe images are annotated with bounding boxes. For the CNN detection model, we produce about 67000 cropped images with IoU < 0.2 as the negative samples and about 42000 cropped positive images with IoU > 0.8. For the DPM detection model, we use about 2000 negative cropped images with IoU < 0.2 and 500 ground truth shoe images as the positive to train a 5-component DPM model. It is worthy to mention that the learnt DPM model is also used to detect the shoe image patches of the third level. The EdgeBox algorithm with default parameters is employed to generate $P = 100$ initial region proposals. Eventually, about 100 shoe images are randomly chosen to produce the ordered pairs for weights learning in RankSVM.

For the evaluation of the retrieval performance, we use 4021 online domain images as the reference gallery and 5021 street domain images as the query where each of them has a counterpart in the reference set.

4.3. Performance Evaluation and Comparison

4.3.1. The Prediction of Clothing Attributes

Our proposed concept of subordinate attribute and SA-CNN network demonstrates its effectiveness on the clothing. We find that the multi-class attribute like "Neckline", "Sleeves Length" and "Category" can be further categorized into subordinate attribute classes by merging attribute values with closer visual appearances. For the attribute type such as "Neckline", we consider "Round Neckline" and "V-Shape Neckline" as a semantically-close SA and the rest as another SA. In the same way, like the attribute "Sleeves Length", the attribute value of "No Sleeve" are treated as one SA

Table 1. Comparison of the multi-class attribute prediction accuracy on clothing attribute dataset

	[1] without pose	CNN	SA-CNN
Neckline	53.2	54.15	**56.50**
Sleeves	67.8	69.0	**71.20**
Category	47.5	53.4	**55.2**

Fig. 5. Examples of clothing attribute prediction results.

and "Short Sleeve", "Long Sleeve" as another SA. For the "Category" attribute, the original 7 semantic attributes can be merged into 3 SA clusters. The performance of the attribute prediction results with our proposed SA-CNN are compared with that using the typical multi-class CNN and the combined traditional hand-crafted features [1] without the side information of pose estimation. Here the comparison is carried out with the results of [1] without pose, because our algorithm only requires the whole image as input without the complicated procedures to discriminate the pose of the given body skeleton. Table 1 demonstrates the comparison results, and we can find that the proposed SA-CNN improves consistently over the CNN by 2.3% for each of the three multi-class attributes. Moreover, the attribute prediction accuracy by SA-CNN outperforms that using traditional hand-crafted features by 3.3%, 3.4% and 7.7%. The improvement for the "Category" is the largest, possibly because the number of misclassification cases between different subordinate clusters is reduced by our proposed SA-CNN. We show three attribute prediction examples in Fig. 5 that are correctly classified (consistent with the ground truth) by the proposed SA-CNN but causes failure to conventional CNN fine-tuned on the clothing dataset.

4.3.2. Cross-domain Shoe Retrieval

We compare the proposed cross-domain shoe retrieval approach with the following baselines using the top-20 retrieval accuracy.

1) Traditional features: GIST feature [4] with 512-dimension and Dense SIFT feature followed by fisher vector

Table 2. Top-20 retrieval accuracy of our system and baseline results

Method	Top-20 Accuracy
Gist feature [4]	11.13
DSIFT + Fisher Vector [5]	20.04
Deep feature (fc6) of Pre-trained CNN	28.28
Deep feature (fc6) of SA-CNN	48.20
Metric Network [6]	52.43
Multi-level feature with SAG-CNN [9]	66.92
Three-level feature (conv5 + fc6) with SA-CNN	**69.11**

encoding (*DSIFT + Fisher Vector*) [5] with the codebook size $D = 64$.

2) Deep feature (fc6) of Pre-trained CNN: deep feature extracted from the fc6 layer of the pre-trained AlexNet [7] with the whole image as the input.

3) Deep feature (fc6) of SA-CNN: deep feature activated from the fc6 layer of the SA-CNN network with the whole image as the input.

4) Metric Network [6]: it is used to evaluate the similarity of feature vectors activated from the fc6 layers of the pre-trained AlexNet on the whole image.

5) Multi-level feature with SAG-CNN [9]: three-level deep feature activated from SAG-CNN.

As we can see from the experimental results shown in Table 2, deep features have a significant improvement compared with that utilizing the traditional features. The proposed multi-level deep features with SA-CNN outperform all the baselines using the deep features from different levels of the network by a large margin. Moreover, compared to a recent work [6], our approach shows a significant enhancement of 14%. A comparison is also conducted with our previous best-performing result in [9], and this work outperforms [9] by about 2%, which validates the advantage of the hierarchical merging of visually similar semantic attributes as the subordinate attribute rather than treating them equally.

5. CONCLUSIONS

We demonstrate a subordinate attribute based convolutional neural network (SA-CNN) for tackling two problems in the fashion field, cross-domain shoe retrieval and clothing attribute prediction. It not only improves the clothing attribute prediction but also represents the appearance of shoe in a discriminative manner. In addition, a discriminative three-level deep feature representation extracted from SA-CNN is presented. We demonstrate that the proposed SA-CNN is effective in the semantic clothing attribute prediction and the activated feature has a significant improvement over other baselines in cross-domain shoe retrieval.

6. REFERENCES

[1] Huizhong Chen, Andrew Gallagher, and Bernd Girod, "Describing clothing by semantic attributes," in *ECCV*, 2012.

[2] John Lafferty, Andrew McCallum, and Fernando Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML*, 2001.

[3] Ran Tao, Arnold WM Smeulders, and Shih-Fu Chang, "Generic instance search and re-identification from one example via attributes and categories," *arXiv preprint arXiv:1605.07104*, 2016.

[4] Aude Oliva and Antonio Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *IJCV*, 2001.

[5] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez, "Aggregating local descriptors into a compact image representation," in *CVPR*, 2010.

[6] M Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C Berg, and Tamara L Berg, "Where to buy it: Matching street clothing photos in online shops," in *ICCV*, 2015.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[8] Pedro Felzenszwalb, David McAllester, and Deva Ramanan, "A discriminatively trained, multiscale, deformable part model," in *CVPR*, 2008.

[9] Huijing Zhan, Boxin Shi, and Alex C. Kot, "Cross-domain shoe retrieval using a three-level deep feature representation," in *ISCAS*, 2017.

[10] Andrea Vedaldi and Karel Lenc, "Matconvnet: Convolutional neural networks for matlab," in *ACM Multimedia*, 2015.

[11] C Lawrence Zitnick and Piotr Dollár, "Edge boxes: Locating object proposals from edges," in *ECCV*, 2014.

[12] Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders, "Selective search for object recognition," *IJCV*, 2013.

[13] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li, "Learning to rank: from pairwise approach to listwise approach," in *ICML*, 2007.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *ECCV*, 2014.