Cross-Domain Shoe Retrieval With a Semantic Hierarchy of Attribute Classification Network

Huijing Zhan, *Student Member, IEEE*, Boxin Shi, *Member, IEEE*, and Alex C. Kot, *Fellow, IEEE*

***Abstract*— Cross-domain shoe image retrieval is a challenging problem, because the query photo from the street domain (daily life scenario) and the reference photo in the online domain (online shop images) have significant visual differences due to the viewpoint and scale variation, self-occlusion, and cluttered background. This paper proposes the semantic hierarchy of attribute convolutional neural network (SHOE-CNN) with a three-level feature representation for discriminative shoe feature expression and efficient retrieval. The SHOE-CNN with its newly designed loss function systematically merges semantic attributes of closer visual appearances to prevent shoe images with the obvious visual differences being confused with each other; the features extracted from image, region, and part levels effectively match the shoe images across different domains. We collect a large-scale shoe data set composed of 14341 street domain and 12652 corresponding online domain images with fine-grained attributes to train our network and evaluate our system. The top-20 retrieval accuracy improves significantly over the solution with the pre-trained CNN features.**

***Index Terms*— Semantic hierarchy, cross-domain, instance retrieval, convolutional neural network.**

I. Introduction

ONLINE shopping grows explosively in recent years. It is reported that more than 100 million Americans purchase online every year and overall income from e-commerce market grows substantially, up to about $350 billion [1]. Fashion products purchase contributes a comparatively high percentage for online shopping; particularly, the commerce on footwear is experiencing an e-boom with revenues from online shoe sales achieving $445.4 million [2] through year 2014-15. Leading shoe manufacturers (such as Addidas, Nike, *etc*.) launch their online shop and provide a variety of customized services, like 3D viewing of shoes, suggesting the right shoe size according to the customers' owned shoes, and recommending the shoes according to your browsing and purchase records. In daily life, such scenarios happen frequently: The shopper may come across his/her favorite shoes in a shop window or notice them worn on others' feet. When the shopper wants to find exactly the same pair from online shoe shops, it is always difficult to describe the target shoes precisely using the text-based search engine. Therefore, visual search for exactly the same shoes from online shops given daily photos, *i.e.*, instance retrieval, is always desired. This problem is called *cross-domain* shoe retrieval: The query and reference images are from different domains. We name the shoe photos captured in the daily life scenario *street domain*, and the photos from online shop pictures *online domain*. Different from the retrieval for other fashion items like clothes, the cross-domain shoe retrieval has its unique challenges from viewpoint variation, self-occlusion, scale variation, and cluttered background, as illustrated in Fig. 1.

Fig. 1. Example street domain and their counterpart online domain shoe images. The street domain photos may have human feet, different poses and viewpoints of shoes, scale variation as well as illumination variation. In contrast, pictures from online shops are captured in fixed viewpoints, ideal illumination conditions with clean background.

The cross-domain shoe retrieval can also be considered as a more challenging type of fine-grained image search problem [3]. Given a query image (*e.g.*, bird), the fine-grained image search aims to find the images belonging to the same subcategory object class as the query, while our cross-domain shoe retrieval is to search the items in a finer manner to return the exact same shoes rather than the same category or style of shoes as the query.

An effective feature descriptor tailored for the object of interest is the key to instance retrieval. Conventional systems focus on comprehensive image representation, and they adopt classic features using global descriptors like GIST [4], color features [5] and local descriptors such as SIFT [6],

Manuscript received December 11, 2016; revised May 30, 2017 and July 17, 2017; accepted July 19, 2017. Date of publication August 4, 2017; date of current version September 21, 2017. This work was supported by the National Research Foundation, Singapore, through the Interactive Digital Media Strategic Research Programme. The work of B. Shi was supported by the project commissioned by the New Energy and Industrial Technology Development Organization. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Wen Gao. *(Corresponding authors: Huijing Zhan; Boxin Shi.)*

H. Zhan is with the Rapid-Rich Object Search Laboratory, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 637553 (e-mail: zh0069ng@e.ntu.edu.sg).

B. Shi is with the Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo 135-0064, Japan (e-mail: boxin.shi@aist.go.jp).

A. C. Kot is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: eackot@ntu.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2017.2736346

LBP [7], *etc*. However, these general image descriptors usually perform poor on objects with large viewpoint variation. Besides viewpoint, the visual discrepancy from street and online domain shoe images caused by self-occlusion, scale variation, and cluttered background could make the feature distance of images depicting the same shoe from different domains even larger than those of different shoes. Existing image-based instance retrieval systems like [8], [9] achieve excellent performances on objects like buildings and logos (mostly planar objects with distinct features), but they cannot be directly applied to shoes partly due to these cross-domain challenges. This motivates us to develop effective feature representations for minimizing the distances of the same shoe in different domains, while maximizing the distances for different shoes.

In fine-grained image search [3] or classification [10], [11], the main challenge lies in how to differentiate the subtle differences (*i.e.*, color of the wing) between subcategories. Examples include identifying different types of birds, species of flowers or breeds of dogs. Due to the descriptive capability of the semantic attribute at capturing the subtle local part differences for fine-grained images, we also incorporate the prediction of the semantic attribute into our retrieval framework. Our goal is to design an appropriate attribute prediction strategy for a more robust and discriminative shoe feature representation for retrieval. Semantic part-aware shoe attributes, *i.e.*, the attributes of key parts of shoes such as toe and heel, encode more discriminative information. However, correct prediction of semantic attributes for shoes is not an easy task, especially in the cross-domain scenario, because semantically close attributes might have extremely similar visual appearance. To make this problem trackable, our key idea is to consider the hierarchy of semantic shoe attributes and focus on the most discriminative attributes by merging attributes with closer visual appearance as *subordinate attribute*. The correct prediction of subordinate attribute is always given higher priority to ensure that shoes with obviously different visual appearances are never confused with each other. We show such an example for "toe" and "heel" in the Fig. 2. Despite a focus on shoes in this paper, it is natural to generalize the concept of the subordinate attribute to other objects with multi-class attribute annotations such as clothes.

We then integrate the Semantic Hierarchy Of attributE and subordinate attribute prediction into the deep Convolutional Neural Network (SHOE-CNN) for discriminative shoe feature learning. 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.

An overview of our cross-domain shoe retrieval system is illustrated in Fig. 3. Given a query shoe image captured in the daily environment (street domain), we first represent it using features extracted from three levels: the whole image (level 1, orange lines and blocks in Fig. 3), the top-3 scored region proposals produced by our proposed region proposal selection approach (level 2, green lines and blocks in Fig. 3), and shoe image patches detected by DPM (level 3, blue lines and blocks in Fig. 3). Then all three-level features are fed forward

Fig. 2. The hierarchical structure of semantic shoe attributes (in black box) and subordinate attributes (SA, in red box) for toes and heels. By grouping "Square Toe", "Round Toe", and "Pointy Toe" as "Closed Toe" (SA1) and "Peep Toe" as "Open Toe" (SA2), the semantic attribute prediction will hardly confuse shoes with visually dissimilar toes. The appearances of the attributes are illustrated in (a). The same applies to heels, as shown in (b).

to SHOE-CNN module, and we use the last convolutional layer (Conv5) and the first fully-connected layer (FC1) from the SHOE-CNN as a three-level visual representation. The final feature vector concatenates these three parts as a comprehensive representation for shoe images. The reference images in the dataset (online domain) follow the similar procedures except that the top-3 scored region proposals are replaced by the whole image, due to their white background. The distances of feature vectors between the query and reference images are compared by the OASIS algorithm. Finally, reference images with large visual and semantic similarities are returned as the retrieval results.

To further deal with the cross-domain challenges from scale variation and cluttered background, we propose a three-level feature representation for shoes as three different scales of input to SHOE-CNN: the first (coarsest) level is the whole image, the second level is the top-3 candidates from a newly proposed region proposal selection algorithm, and the third (finest) level is the detected shoe parts by Deformable Part Model (DPM) [12]. Features activated from the three-levels of SHOE-CNN are integrated to encode both global structures and local patch details, which capture the appearances of shoe images in a coarse-to-fine manner.

Finally, similarity metric learning algorithm using triplet ranking objective is applied to handle with the visual differences of the same shoe between the query image and reference images from different domains.

The major contributions of our work are threefold:

- We investigate the semantic hierarchy of attributes and propose the subordinate attribute to ensure that the

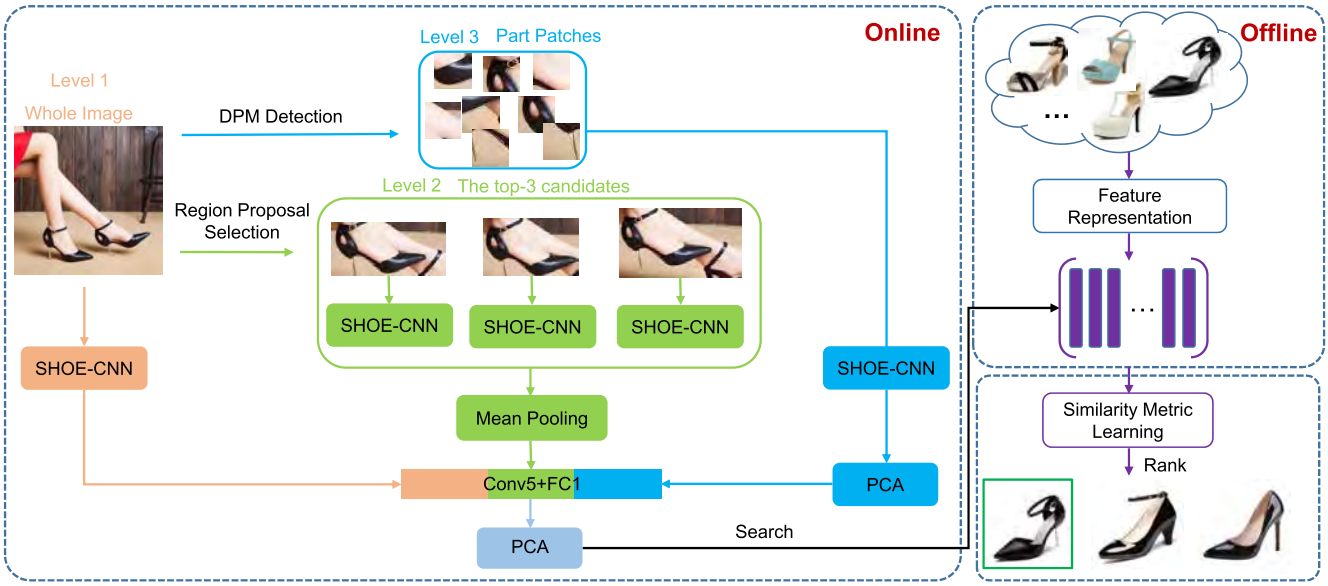


Fig. 3. The overview of our proposed cross-domain exact shoe retrieval system (best viewed in color).

dominant visual dissimilarity on shoe images is more reliably retained and design a novel loss function for SHOE-CNN to benefit the effective feature learning. To the best of our knowledge, we are the first to explore the intrinsic hierarchical properties of semantic attributes of shoes for cross-domain instance retrieval.

- We develop a three-level feature representation to effectively extract the multi-scale information from shoe images and combine them with the multi-layer features of SHOE-CNN to conquer the unique challenges of cross-domain shoe retrieval.
- A large-scale shoe dataset, composed of 14341 street domain and 12652 corresponding online domain images with fine-grained attributes, is collected to verify the effectiveness of our shoe retrieval system.

II. Related Works

In this section, we briefly introduce the related works, including instance retrieval, attribute-based representation in image retrieval, and image retrieval for fashion items, in particular for shoes.

A. Instance Retrieval

Given a query image depicting a particular object, instance retrieval aims at finding the images containing exactly the same object in the reference database. Recent works on instance search approaches focus on either designing a robust image representation (feature-based) or developing the metric learning technique to evaluate the similarity between the query and reference images (metric learning based).

The state-of-the-art feature-based methods represent the images by bags of local features [6] followed by feature encoding approaches such as VLAD or Fisher Vector [13] that forms a more compact feature representation. However, it loses the robustness for 3D objects like shoes due to the large viewpoint variation. In recent years, CNNs have greatly improve the state-of-the-art performance in various computer vision tasks such as image classification [14], image segmentation [15], [16], object detection [17], *etc*. CNN-based features also show their excellence in instance retrieval. In [18], neural codes, a high-level descriptor invoked from the upper-layer of CNN network, show advanced performance. Razavian *et al.* [19] suggested that features activated from the intermediate layers (convolutional layers) together with Spatial Pyramid Pooling (SPP) [20] can be used to better preserve the spatial information. Moreover, combining both features extracted from the fully-connected and last convolutional layers could further enhance the retrieval performance [21]. The multi-scale CNN features are also beneficial for image representation. Gong *et al.* [22] combined activations extracted from local image windows at multiple scale levels and achieved excellent results in scene retrieval. Yao *et al.* [23] developed a multi-level feature description to capture the coarse-to-fine visual characteristics of the fine-grained object categories.

A variety of similarity learning approaches have been proposed to facilitate the design of an effective instance retrieval system. The triplet-based learning approach [24]–[27] is widely used for image similarity metric modeling, with one image denoted as the anchor, a positive image similar with the anchor and a dissimilar image with the anchor. Chechik *et al.* [25] developed an Online Algorithm for Scalable Image Similarity Learning (OASIS) to learn a bilinear image similarity metric over LBP and color histograms using triplet based data. Wu *et al.* [28] leveraged the deep learning network to learn a ranking model with the hand-crafted features as input. Wang *et al.* [26] proposed a novel multi-scale CNN network structure to learn a deep ranking model for measuring the image similarities directly from the pixels.

Although the CNN-based ranking model has achieved promising results, it requires a large amount of training data and the final ranking accuracy highly depends on how the triplets are sampled. Choosing the triplets inappropriately might lead to the slow convergence speed and local minima [29]. In this paper, we leverage OASIS algorithm to learn the shoe similarity metric on top of features activated from our proposed SHOE-CNN network. Our system focuses more on designing effective features for representing shoes. We adopt the multi-scale information of shoes by depicting a three-level feature representation.

B. Attribute-Based Representation

Due to the descriptive capability of the attribute-based representation to cross the boundary of different object categories, it has been widely used to solve a variety of object recognition and classification problems [30]–[34]. The task of the fine-grained classification greatly benefits from the introduction of the fine-grained attribute into the local part, which captures the subtle visual differences between different subcategories of a basic category (*e.g.*, bird [32], flower [33], aircraft [34]). Vedaldi *et al.* [34] investigated the influence of part detection and semantic attributes on understanding the fine-grained aircraft in detail. Berg and Belhumeur [32] proposed Part-based One-vs-One Features (POOFs) which extracted local features from the semantic parts of the objects for fine-grained classification and attribute estimation. Semantic attributes were also demonstrated to be effective in the fine-grained image retrieval problem. Li *et al.* [35] integrated the part-aware shoe attributes for fine-grained sketched-based image retrieval. The performance of the clothing retrieval also benefited from the incorporation of clothing attributes and aligned human parts [36].

Recent works incorporated attribute classification in the CNN learning process for both achieving a better attribute prediction accuracy and generating a more discriminative feature. Zhang *et al.* [37] developed a tree-structured poselet-specific CNN for human attribute prediction and further proposed an effective pose-normalization representation with features extracted from the poselet-specific CNN. However, their proposed CNN was used to predict the binary attributes. Huang *et al.* [38] integrated tree-structured CNN with fine-grained clothing attribute labels in learning effective clothing features for retrieval. However, they merely utilized the attribute prediction as a regularizer for clothing retrieval, the inherent correlation of the multi-class attribute is not taken into account.

There are also existing works investigating the coherence of the attributes to further improve the attribute prediction accuracy. Chen *et al.* [39] captured the mutual dependencies between the clothes attributes with the Conditional Random Field (CRF) model. Wang and Mori [40] proposed to use a tree-structured graph to model the relationship among various attributes. However, to the best of our knowledge, the inherent properties of the multi-class attribute have not been explored for the semantic attributes of shoes. We further organize semantic attributes in a hierarchical manner as subordinate attributes in the CNN attribute classification framework for a more discriminative representation.

C. Image Retrieval for Fashion Items

Existing works on fashion items mainly target on clothing, like clothes parsing [41], [42], clothing retrieval [36], [43], [44], branded handbag recognition [45], *etc*. There exists several works addressing the domain-gap in the clothing retrieval problem. Liu *et al.* [44] developed a multi-task CNN that jointly predicted the attribute and landmark location. The features activated from the learnt CNN were used to perform the clothing retrieval. Chen *et al.* [46] developed a deep domain adaption network to model the clothing features extracted from the different domains simultaneously keeping the consistency of matched pairs between different domains. Zhou *et al.* [43] proposed a hybrid topic model to bridge the gap between semantic text descriptions and extracted low-level features. In contrast, the studies on image-based shoe retrieval is still at infancy stage. Kovashka *et al.* [47] developed a shoe recommendation system utilizing relative attribute based pairwise comparison. Huang *et al.* [48] proposed a similar shoe retrieval framework by integrating attributes in both the part detection and shoe retrieval step, with the help from manual attribute and part annotation. Kiapour *et al.* [49] focused on the exact street-to-shop product retrieval by adopting a category-independent metric network to evaluate the similarity of image pairs. The location of the query object is provided as a prior knowledge. However, we propose a region proposal selection approach to localize the shoes and the top-3 selected region proposals are regarded as one level of the feature of our system. We represent shoe features using a three-level structure combined with semantic hierarchy of attribute-based CNN, which facilitates us to learn the semantic-aware features for cross-domain exactly the same shoe retrieval. We also focus on the hierarchical properties of the semantic shoe attributes, and propose the subordinate attribute to improve the prediction based on these discriminative attributes.

III. Our Approach

In this section, we first present the proposed SHOE-CNN which takes the hierarchical properties of multi-class attribute into consideration and aims to reduce the prediction errors between the attributes that are far away in appearances. Then, based on the SHOE-CNN, we develop the three-level deep shoe feature representation that captures the appearances of shoes in the global aspects and local details.

A. SHOE-CNN: Semantic Hierarchy of attributE Based Convolutional Neural Network

As illustrated in Fig. 2, treating different semantic attributes equally loses the discriminative visual features for shoe parts. We utilize the hierarchical structure of semantic attribute grouping and propose a corresponding CNN with novel loss function (SHOE-CNN) to address this issue.

1) Network Architecture: The network architecture is shown in Fig. 4. Insipred by the work of Zhang *et al.* [37] which designed attribute-sharing deep neural network for the human attribute prediction, we make appropriate adaption

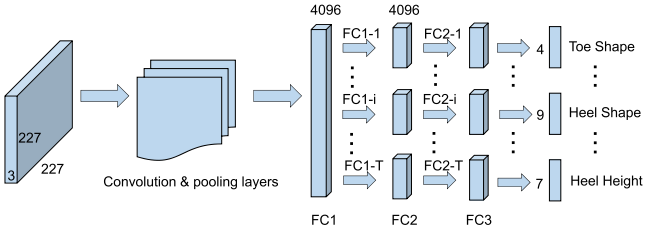


Fig. 4. The architecture of our SHOE-CNN network. It consists of five convolution layers and one fully-connected layers shared for all the attributes. However, FC1 branches out T attribute-specific layers (T is the number of semantic shoe attribute types) thus each attribute owns independent FC2 and FC3 layers.

based on the AlexNet [14] by adding a tree-structure set of fully-connected layers after FC1 layer to obtain attribute-sensitive features. Our network contains five convolutional layers and one fully-connected layer with shared parameters for all the attributes, while the last two fully-connected layers are not shared. For each attribute, its last fully-connected layer branches out several units, according to the number of potential sub-class attributes values. For example, the part-aware attribute "toe shape" takes 4 values so its specific fully-connected layers branches out 4 units, indicating the probabilities of having the corresponding attribute value. We adopt such a structure because different attribute predictions require different features for prediction; for example, detecting whether the toe shape or the heel shape of a particular shoe requires different features.

2) Loss Function: Our attributes mostly belong to the multi-class attribute type (*i.e.*, take multiple real values), 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 $x_i \in \{1, 2, \ldots, N\}$ with $y_i \in \{1, 2, \ldots, K\}$ 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 sample. For the proposed subordinate attribute, the K values can be further merged into M distinct values. For the toe shape attribute, it originally consists of four attribute values ($K = 4$), corresponding to "Square Toe", "Round Toe", "Pointy Toe" and "Peep Toe". Because the appearances of the first three sub-class attributes obviously have physically-closed toe structure, we consider them to be semantically close and term them as one subordinate attribute (SA1 in Fig. 2(a)). Meanwhile, the peep toe has an open shape thus we term it as another subordinate attribute (SA2 in Fig. 2(a)). Therefore, as to the toe shape attribute, the previous four semantic attribute values ($K = 4$) are further clustered as two subordinate attribute classes ($M = 2$). Our key idea is to penalize the mismatching between subordinate attributes by a larger weight, so that the misclassification between "Round Toe" and "Peep Toe" is imposed with larger penalty than that between "Round Toe" and "Pointy Toe". Let $h_j^{(i)} (j = \{1, 2, \ldots, K\})$ indicate the activation value of the node j from the last fully-connected layer (FC3) in Fig. 4, 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 case, the classification error made within each subordinate attribute cluster is smaller than the errors between clusters. Thus, the penalty between subordinate class 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. Here assume that we have the K attribute values which can be further categorized into M subordinate attribute classes. Let $\{\mathbb{G}_1, \mathbb{G}_2, \ldots, \mathbb{G}_M\}$ represent M semantically-close clusters, where M is the number of subordinate attribute classes (clusters). Here $\mathbb{G}_t$ $(t = \{1, 2, \ldots, M\})$ 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:

$$J = J_0 + J_1, \tag{3}$$

with

$$J_1 = -\frac{1}{N}\left[\sum_{i=1}^{N} \lambda\left(\sum_{j\in\mathbb{G}_1} \delta\ (y^{(i)} = j) \log P_{\mathbb{G}_1}^{(i)} + \sum_{j\in\mathbb{G}_2} \delta\ (y^{(i)} = j) \log P_{\mathbb{G}_2}^{(i)} \ldots + \sum_{j\in\mathbb{G}_M} \delta\ (y^{(i)} = j) \log P_{\mathbb{G}_M}^{(i)}\right)\right]. \tag{4}$$

From Eq. 4 we can see 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:

$$P_{\mathbb{G}_t} = \sum_{j\in\mathbb{G}_t} P_j^{(i)}, \tag{5}$$

where λ is a regularization parameter controlling the balance between the original loss term J_0 and the subordinate attribute classification term J_1. As we can see if we set λ as a small values, 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 the new loss term J_1 with respect to the output of the last fully-connected layer $(h_j^{(i)}; j = \{1, 2, \ldots, K\})$ is computed and the back-propagation is applied to optimize the parameters for the network.

The partial derivatives of J_0 is provided in the literature [50], as demonstrated in Eq. 6.

$$\frac{\partial J_0}{\partial h_j^{(i)}} = \frac{1}{N}\Big[P_j^{(i)} - \delta(y^{(i)} = j)\Big], \tag{6}$$

Our focus is then to obtain the derivatives of the second term J_1 with respect to the output of the last fully-connected layer. Different from the partial derivative obtained with respect to the original softmax loss function J_0, the partial derivative of the term J_1 with respect to the activation value of FC3 layer is different according to the different input. Assume that $y^{(i)} \in \mathbb{G}_t$, the deduction procedure of loss J is shown below:

$$\frac{\partial J}{\partial h_{j\notin\mathbb{G}_t}^{(i)}} = \frac{1}{N}\Big[(\lambda+1)P_j^{(i)} - \delta(y^{(i)} = j)\Big], \tag{7}$$

and

$$\begin{aligned}\frac{\partial J}{\partial h_{j\in\mathbb{G}_t}^{(i)}} &= \frac{1}{N}\left[-\lambda\left(\frac{P_j^{(i)}}{P_{\mathbb{G}_t}} - P_j^{(i)}\right) + P_j^{(i)} - \delta(y^{(i)} = j)\right] \\ &= \frac{1}{N}\left[P_j^{(i)}\left(\lambda + 1 - \frac{\lambda}{P_{\mathbb{G}_t}}\right) - \delta(y^{(i)} = j)\right].\end{aligned} \tag{8}$$

Then the generalized partial derivatives of Eq. 3 is further formulated as:

$$\begin{aligned}\frac{\partial J}{\partial h_j^{(i)}} = \frac{1}{N}\Bigg\{&\delta(j \notin \mathbb{G}_t)\Big[(\lambda+1)P_j^{(i)} - \delta(y^{(i)} = j)\Big] \\ &+ \delta(j \in \mathbb{G}_t)\Big[(\lambda + 1 - \frac{\lambda}{P_{\mathbb{G}_t}})P_j^{(i)} - \delta(y^{(i)} = j)\Big]\Bigg\}.\end{aligned} \tag{9}$$

B. Three-Level Feature Representations for Shoes

To address the challenges of scale variation and background clutter, we design a three-level feature representation for shoe images. As illustrated in Fig. 3, the features are a combination of SHOE-CNN activated deep features from the whole image (Level 1), the top-3 scored region proposals (Level 2) with large likelihood to contain the shoe, and the fine-grained local part patches detected by DPM (Level 3). Obtaining the first and third level features are straightforward, so we focus on introducing our proposed region proposal selection approach for second level representation.

The goal of generating object region proposals is to produce a set of candidate windows for the subsequent detection process [51]. The initially generated region proposals contain a large number of low-quality candidate boxes, with low Intersection-Over-Union (IoU) scores for the annotated object. IoU is calculated by the intersection of the candidate bounding boxes with the ground truth box divided by the union of them. Thus, it is essential to develop a ranking strategy to rank the list of produced region proposals and determine the location of the object based on the top-ranked candidate boxes. We integrate three criteria for evaluating the quality of the region proposals: 1) the CNN probability score, 2) the confidence scores produced by EdgeBox, and 3) DPM detection score are employed.

1) *Confidence Score by EdgeBox:* We adopt the Edgebox [51] algorithm to generate the initial region proposals for the subsequent proposal selection procedure. The confidence score returned by EdgeBox is denoted as e.
2) *Probability Score of CNN Model:* We use deep CNN as a binary classifier to differentiate whether an image patch belongs to the foreground or the background. The activation of the softmax layer is extracted to encode the probability of the region proposal being a shoe region, denoted as c.
3) *Confidence Score of DPM:* With the ground truth box provided in the training stage, a deformable part model (DPM) model is learned to detect the location of the shoe. The DPM returns a quality score computing the overlap of its detection bounding box with the region proposals, denoted as d.

After obtaining the three types of scores e, c, and d, we apply rankSVM [52] to learn the weights balancing the contributions of these three scores, and then the object proposals are ranked according to the combined score. The details are as follows:

For a query image in the street domain, we first measure the IoU score of each candidate region proposal with the annotated shoe. Let u_i denote the IoU score of the i-th region proposal, which is represented as a vector $\mathbf{h}_i \in \mathbb{R}^3$ by concatenating e_i, c_i and d_i. Thus an image can be considered as the bag of confidence feature $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_P]$. As rankSVM is based on pairwise comparison, the generated P region proposals (generated by each image) have $P \times (P-1)/2$ pairs. Among them, we randomly sample M pairs of region proposals denoted as $\mathbb{O} = \{(s_k, t_k), k = \{1, 2, \ldots, M\}\}$ and calculate their pairwise relevance label $T(s_k, t_k)$. We define $T(s_k, t_k) = 1$ if $u_{s_k} > u_{t_k}$; otherwise, $T(s_k, t_k) = -1$.

With the supervised pairwise information, our goal is to learn a ranking function $f(\mathbf{h}) = \mathbf{w}^\top \mathbf{h}$ which predicts a score for each region proposal $\mathbf{h}$. Moreover, the function $f(\cdot)$ is capable of estimating the relevant relationship between data pairs with the following constraint:

$$\forall u_{s_k} > u_{t_k} : f(\mathbf{h}_{s_k}) > f(\mathbf{h}_{t_k}), \tag{10}$$

Then the rankSVM model is built by minimizing the objective function:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{(s_k,t_k)\in\mathbb{O}} l(\mathbf{w}^\top \mathbf{h}_{s_k} - \mathbf{w}^\top \mathbf{h}_{t_k}). \tag{11}$$

where l is a loss function with the form $l(t) = \max(0, 1-t)$ and C is the trade-off parameter. The learnt $\mathbf{w}$ indicates the relative importance of the confidence scores from three conventional detection models mentioned above. The final concatenated quality score for each image is denoted as $\mathbf{J} = \mathbf{w}^\top \mathbf{H} = \mathbf{w}^\top [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_P]$, where $\mathbf{J}$ is P-dimensional vector, with each element indicating the quality score for its corresponding region proposal. The top-3 scored region proposals are chosen as the final detection results. Fig. 5 illustrates the region proposal selection results.

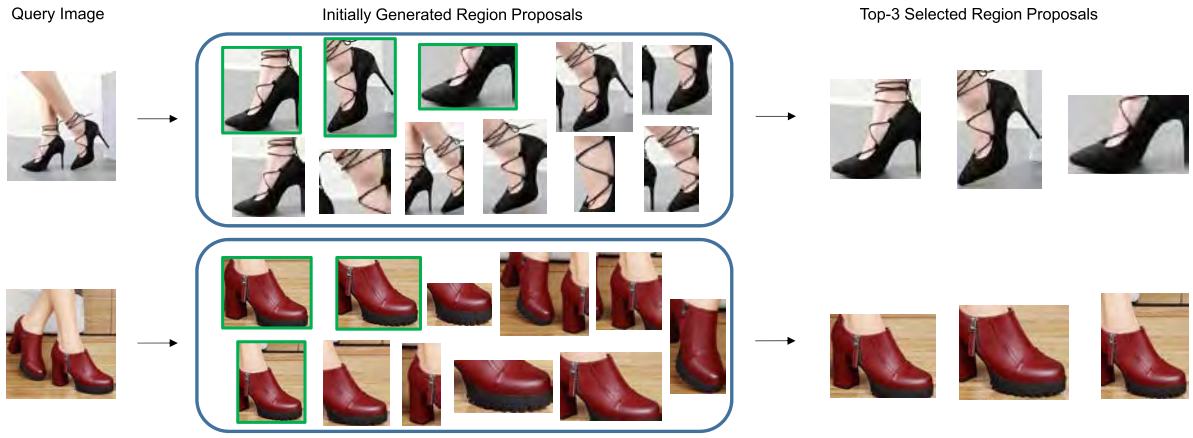


Fig. 5. The 1st column pictures indicate the query real-world shoe photos and the rightmost three columns illustrate the top-3 region proposals selected by our proposed method. The middle part shows the initial coarse region proposals generated by the EdgeBox algorithm [51], and the top-3 scored candidates are highlighted in green.

C. Deep Features Activated From Three-Level SHOE-CNN

The feature representation for shoes is composed of SHOE-CNN activated features from the above mentioned three scale levels. The deep features activated from FC1 layer is supposed to convey the overall content of the image, however it fails to describe the appearance of local image patches. To overcome its lack of spatial information, we integrate the Conv5 feature and the FC1 feature to describe the appearance of the image in both global and local aspects.

Given an input image, the image is first resized to a fixed size defined by AlexNet and fed forward to the network. The features are extracted from the last convolutional layer (Conv5) and we apply the 2-level pyramid mean-pooling [20] $(2\times 2, 1\times 1)$ on the Conv5 feature map with the size (6×6). Two feature outputs are generated with the size, which are equal to $2\times 2\times 256 = 1024$ and $1\times 1\times 256 = 256$, respectively. Eventually, we obtain a $1024 + 256 = 1280$-dimensional Conv5 layer feature representation after spatial pyramid pooling (SPP). The deep feature based representation for an image is obtained by combining the activated features from the FC1 layer and the Conv5 layer, thus generating a feature vector with dimension of $1280 + 4096 = 5376$.

The final feature representation for shoes consists of the visual feature-based representation from three levels, so the resulting feature has a dimension of $5376 + 5376 + 1024 = 11776$, which is further reduced to 1024-D. Therefore, each query image is described by a 1024-D feature vector after L2 normalization.

D. Similarity Metric Learning

Although the cosine similarity is a popular metric to evaluate the relevance between the query and reference images in the content-based image retrieval, it is not well-suited for the cross-domain shoe retrieval problem because the cosine similarity between low-level visual feature vectors cannot reflect semantic relevance of the pictures from two domains.

To evaluate the similarities between images from different domains, we use metric learning to learn the similarity measure by the image triplets. An image triplet typically consists of one reference image and another two photos, with one similar to the reference image in terms of category-level label or instance-level object ID label (also indicated as the relevant class) and the other one shares a different label with the reference (indicated as the irrelevant class). In our case, each shoe image is associated with a unique product ID. Thus we can employ the shoe ID label to identify the relevant and irrelevant class indicated as positive and negative samples. With the available triplet-based side information, the OASIS algorithm [25] is utilized to evaluate the similarity between pairs of shoes on top of the SHOE-CNN activated deep features. We choose to use OASIS because it is the state-of-the-art approach [53] for learning the semantic similarity, which employs the large margin criterion that enables fast training convergence [25] without degrading the performance.

To make this paper self-contained, we briefly introduce how OASIS algorithm is applied to solve our problem. Given a set of images denoted as $\mathbb{P}$, each image can be represented by a vector $p_i \in \mathbb{R}^d$, where $d = 1024$. We assume that for each image p_i, its relevant images sharing the same object ID and the dissimilar object instances with a different ID number can be represented as p_i^+ and p_i^-. With the triplets of images (p_i, p_i^+, p_i^-), the core of similarity measurement task is to learn a mapping function S for the similarity relevance between relevant pairs (p_i, p_i^+) is larger than irrelevant pairs (p_i, p_i^-). The similarity function S parameterized by W is formulated as:

$$S_{\mathbf{W}}(p_i, p_j) = p_i^T \mathbf{W} p_j, \tag{12}$$

where $p_i \in \mathbb{R}^d$ and $p_j \in \mathbb{R}^d$ can be any two feature vectors of images in the set $\mathbb{P}$.

We choose p_i as any sampled street domain images from the image set, and both of the images p_i^+ and p_i^- are obtained from the online shoe shop domain shoe pictures. In an image triplet of shoes, (p_i, p_i^+) is the matched pairs, while (p_i, p_i^-) is the un-matched pairs between the street domain and the online product shoe photo domain. We need to minimize the

Fig. 6. The Image collage from our complete dataset: Left picture demonstrates the images from the street domain and right picture shows the photos from the online domain. Example corresponding images of exact the same pair of shoes from both domains are illustrated in close-up views. Each pair of shoes are represented using three images from the street domain (highlighted in red boxes) and its counterpart online shoe images (highlighted in green boxes). Please zoom-in the electronic version for better details.

distance of feature vectors of the same shoe from different domains, so we define a loss function for each of the triplet as

$$l^t_{\mathbf{W}}(p_i, p_i^+, p_i^-) = \max\{0, \gamma - S_{\mathbf{W}}(p_i, p_i^+) + S_{\mathbf{W}}(p_i, p_i^-)\}, \tag{13}$$

where γ is the margin constant. The loss function penalizes the positive matched pairs that are far apart and dissimilar unmatched pairs with close distance. Our goal is to minimize a global loss $L_{\mathbf{W}}$ that accumulates the hinge loss over all the triplets in the training stage:

$$L_{\mathbf{W}} = \sum_{(p_i, p_i^+, p_i^-) \in \mathbb{P}} l^t_{\mathbf{W}}(p_i, p_i^+, p_i^-). \tag{14}$$

Passive-Aggressive algorithm [54] provides a solution to minimize the loss over $L_{\mathbf{W}}$ and obtain the optimized $\mathbf{W}$.

Finally, for each query image $q_i \in \mathbb{Q}$, we calculate its similarity score with the product shoe images in the reference set and rank the images according to the similarity score. We show the top-20 images in the retrieval list and expect that the exact same shoes from the online-shop domain appear in the top positions.

IV. Dataset Construction

To the best of our knowledge, the UT Zappos50K [40] is the only publicly available shoe image dataset. It consists of about 50,000 catalog images collected from *Zappos*,[1] which covers 4 major categories. Their images are captured in the left-side viewpoint with the white background. This dataset cannot be used directly for our cross-domain shoe retrieval problem, because 1) they do not provide sufficient corresponding images from the street domain, and 2) they only provide 4 types of attributes which are not enough to accurately describe the appearances of shoes, and 3) the ground truth attribute labels are provided in the form of relative attribute strength comparison based on the image pairs (*e.g.*, with respect to the relative attribute "comfort", shoe A is more comfortable than shoe B). So we need to build a new dataset to train and evaluate our cross-domain shoe retrieval system.

We collected 14341 daily shoe photos of about 5400 items from the street domain and 12652 shoe pictures from the online domain with fixed pose in the left side 45 degree view and clean background. The images are crawled from two large online shopping websites, *Jingdong*[2] and *Amazon*.[3] More specifically, with respect to the online domain images, 4907 images are obtained from *Jingdong* while the remaining are obtained from *Amazon*. For images from *Jingdong*, each shoe item has several street domain photos and its corresponding online shoe photos with clean background; for images Amazon, they only provide 5-7 shoe images with white background and we only take the one with left side 45 degree view. The complete dataset [4] and example corresponding shoe images from two different domains are illustrated using the collage and close-up example images in Fig. 6.

The shoe photos from *Jingdong* are provided with fine-grained attribute annotations, however, the images crawled from *Amazon* don't have such well-defined attributes. Our attribute annotation is initialized from an automatic extraction of the text descriptions next to the crawled images. We then

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

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

[3] https://www.amazon.com/

[4] Available for download at: https://sites.google.com/view/crossshoe/home

TABLE I
EXAMPLE ATTRIBUTE VALUES IN OUR DATASET. THE NUMBERS IN PARENTHESIS INDICATE THE TOTAL NUMBER OF VALUES FOR THE CORRESPONDING ATTRIBUTE TYPES

Attribute Name	Attribute Sub-class Values (Total Number)
Close Style	Line-strap, tie, zip...(9)
Toe Shape	Round Toe, Pointy Toe...(4)
Heel Shape	Platform, Flat, Block...(8)
Color	Red, White, Green...(20)
Patterns	Cartoon, Multi-Color, Pure-Color...(13)
Elements	Platform, Zip, Knots, Rivet, Ribbon,...(31)
Styles	Simple, Academic, Roman-Style...(11)
Types	Low-Cut, High Boot...(8)
Suitable Age	Old, Young, Mid-Age, Child(5)
Shoe Width	Low-Cut, Mid-Cut, Deep-Mouth...(4)
Heel Height	Super-High, Middle-High, Flat,...(7)

manually re-touch the attributes using the following three steps: 1) remove attribute types that are not visible from the image such as materials of the sole and shoe inner material types, *etc*; 2) re-annotate crucial attribute labels like heel type, toe shape and color, *etc*; 3) merge the attribute values that are semantically close and visually too difficult to differentiate, for instance, "beige" and "ivory teeth" can be considered as one attribute named "white". Finally, we obtain 11 types of semantic attributes with 120 values in total. The specified attribute categories and their corresponding example attribute values are shown in Table. I.

To facilitate the process of training and evaluating our proposed region proposal selection model, a subset of shoe images in our dataset (about 2000 daily shoe photos) are annotated with bounding boxes. In the case of a pair of shoes shown in the image, each single shoe is enclosed by a bounding box. Therefore, to our knowledge extend, our dataset is the first large-scale shoe image dataset providing street domain and online domain matching image pairs with semantic annotations.

V. EXPERIMENTAL RESULTS

A. Data Preparation and Implementation Details

As illustrated in Fig. 3, the representation for a query image consists of SHOE-CNN activated features with images/patches of three levels. In the following, we first introduce how we split the data for training and evaluating the SHOE-CNN attribute classification model. Then we mention about the data used for training the corresponding models in generating the three-level images/patches. The first level image is not referred, which is the query image. Finally, the data used in the retrieval stage is introduced.

We train a different SHOE-CNN attribute classification model for each domain respectively because the appearance of the shoe attribute shows more variance in the street domain than in the online domain, due to the viewpoint change, *etc*. For training the SHOE-CNN model in the street domain, we generate about 22500 cropped images (top-3 scored region proposals from 7500 daily photos) of about 2000 items using our proposed region proposal selection approach. The remaining 6841 images are used for attribute evaluation. While for the SHOE-CNN training in the online domain, 1800 clean shoe photos are fed forward to network for training and the remaining images from *Jingdong* are used for evaluating the attribute classification performance of the model in the online domain. The number of training images is further increased to 28800 images after simple data augmentation technique like cropping extracted from four corners and center of the image, rotation as well as horizontal flipping.

The training and evaluation of the SHOE-CNN are conducted using the MatConvNet toolbox [50]. The convolutional layers of SHOE-CNN are initialized from the AlexNet model [14]. For the tree-structured layers of SHOE-CNN, the weights are initialized from zero-mean Gaussian distribution. The learning rate is set to 0.001 and divided by 10 after 10 epoches. The batch size is set to 256 for each epoch and the SHOE-CNN for both domains converge after about 20 epoches.

The second level images (top-3 scored region proposals) are obtained based on the confidence scores of three models corresponding to EdgeBox [51], CNN detection model and DPM detection model [12]. The RankSVM model is employed to learn the weights. The training of each detection model and the rankSVM are introduced below:

1) *EdgeBox:* The EdgeBox algorithm with default parameters [51] are used to generate $P = 100$ initial region proposals and the confidence scores.
2) *CNN Detection Model:* We sample image patches randomly around the ground truth bounding box regions. For the positive images, the cropping is done such that IoU > 0.8. While the negative samples are cropped with IoU < 0.2. Around 800 images are used in the CNN detection model training phase. For each image, 50 positive and 80 negative cropped images are generated. Thus, we have about 40000 positive and 64000 negative training examples in total.
3) *DPM Detection Model:* The positive images are ground truth of annotated shoe images and the negatives images are sampled from the initially generated region proposals with IoU < 0.2 with the ground truth of annotated bounding boxes. About 500 positive shoe images and 2000 negative shoe images are sampled from our dataset, with which we train a 5-component shoe model.
4) *RankSVM Training:* We randomly choose 200 shoe images to generate the ordered pairs for weight learning. About $M = 2000$ region proposal pairs are randomly sampled for each shoe image. To evaluate the performance of region proposal selection, the rest shoe images with annotated bounding box are used to evaluate the region proposal selection performance. C is set to 0.001 based on the validation set.

For the retrieval experiment, the dataset is divided into two parts. The first part is used for the similarity metric learning, which contains about 5400 daily shoe photos (randomly sampled) and their counterpart online shoe images. The rest are used to evaluate the retrieval performance, 5021 daily shoe photos are used as the query with each one having the exact

TABLE II

ATTRIBUTE PREDICTION OF THE PROPOSED SHOE-CNN COMPARED WITH OTHER APPROACHES AND DIFFERENT SETTINGS OF SUBORDINATE ATTRIBUTES. (a) STREET DOMAIN. (b) ONLINE DOMAIN.

(a)

Attribute Name	Chen *et al.* [39]	Ozeki *et al.* [55]	CNN	SHOE-CNN			
				$M = 2$	$M = 3$	$M = 4$	Ours
Close Style	55.10	57.97	60.29	60.10	60.42	60.40	**60.58**
Toe Shape	57.96	61.72	68.21	69.03	68.62	68.50	**69.33**
Heel Shape	60.49	60.33	62.38	62.47	**63.75**	63.49	63.36
Patterns	74.60	76.43	78.02	78.04	78.85	78.16	**79.75**
Elements	26.92	27.94	27.61	27.63	28.33	28.36	**29.11**
Styles	30.96	35.73	**38.12**	35.20	35.74	35.64	35.78
Types	51.06	47.07	52.13	53.18	53.35	53.30	**55.35**
Suitable Age	82.09	83.54	84.05	84.88	83.85	82.45	**85.13**
Shoe Width	65.45	66.78	67.62	68.28	68.88	68.57	**69.29**
Heel Height	52.17	50.19	57.17	57.32	57.51	57.89	**58.03**
Color	55.08	64.10	67.80	68.43	68.93	68.54	**70.17**

(b)

Attribute Name	Chen *et al.* [39]	Ozeki *et al.* [55]	CNN	SHOE-CNN			
				$M = 2$	$M = 3$	$M = 4$	Ours
Close Style	64.64	71.65	73.04	72.95	71.76	72.56	**73.90**
Toe Shape	71.00	80.48	81.11	82.45	81.63	81.24	**84.32**
Heel Shape	70.19	73.68	74.32	75.22	77.31	75.17	**79.26**
Patterns	83.57	83.52	84.91	84.15	84.05	84.86	**85.10**
Elements	28.24	32.44	34.79	34.28	35.03	34.88	**35.26**
Styles	32.08	35.77	38.02	38.21	38.06	**38.72**	38.35
Types	52.77	58.12	**62.08**	58.38	59.04	59.80	60.42
Suitable Age	90.31	88.70	**90.89**	89.70	90.22	89.84	90.84
Shoe Width	68.76	71.20	72.62	72.76	73.61	73.37	**74.56**
Heel Height	62.74	64.57	65.40	66.85	67.32	67.69	**67.92**
Color	66.72	78.47	78.55	81.67	81.10	81.15	**82.06**

same shoe in the reference set and the remaining online shoe images are used as the reference gallery.

B. Evaluations of the Proposed System

1) Attribute Prediction Using SHOE-CNN: The attribute prediction performance is evaluated based on the mean average precision (MAP). Table. II and Table. II demonstrate the results of each method in the street domain and online domain, respectively. Here Chen *et al* [39] utilized the combination of traditional golden features (*e.g.*, SIFT, color in the LAB space, *etc*) and CRF model for attribute prediction. Ozeki and Okatani [55] and CNN are deep-based methods without incorporating the hierarchical properties of subordinate attribute in the attribute prediction. On average the attribute prediction using the SHOE-CNN (Ours) is 1.2% higher than the tradition CNN in the street domain and about 1.4% higher in the online domain, which demonstrate the effectiveness of subordinate attributes. We find that the proposed SHOE-CNN with our defined subordinate attributes (denoted as Ours in Table II) outperform other compared methods in most of the attributes, except the attribute like "Suitable Age" and "Types" in the online domain, where the attribute prediction accuracy using the simple CNN slightly outperforms SHOE-CNN. It is mostly due to the data distribution is relatively biased in these two categories of attributes. For example, for the attribute "Suitable Age", the number of images having the attribute value "Youth" is about 9 times more than that of the second largest attribute value "Middle Age".

To further evaluate the effectiveness of the defined subordinate attribute, we vary the number of subordinate attribute M (in Section III-A.2) for each part-aware multi-class attribute as follows: 1) two subordinate attributes ($M = 2$): This is based on whether the given image has a particular property or not. For example, we just identify that the toe of the shoe is open or closed rather than consider the specific shape of it; and for the "Heel Shape", we just consider whether the sole is flat or not. In this way, all the part-aware multi-class attributes are clustered into two SAs. 2) three subordinate attributes ($M = 3$): For the above two subordinate attributes, we further partition the larger SA subset into two SAs. Like the "Toe Shape", the original SA, composed of "Square Toe", "Round Toe" and "Pointy Toe", can be further classified into two SAs. One indicates the non-pointy toe, corresponding to "Square Toe" and "Round Toe"; while the other SA refers to "Pointy Toe". 3) four subordinate attributes ($M = 4$): Both the two SAs obtained from 1) are further partitioned, thus the particular part-aware multi-class attribute is classified into four SAs. The results shown in Table II demonstrate that our definition of subordinate attributes outperform other settings of subordinate attributes. The complete definition of subordinate attribute groupings with different values of M as well as our definition are provided in the supplementary material.

2) Evaluation of the Generality of the SHOE-CNN: We perform experiments using the clothing dataset [39] with semantic attribute labels to evaluate the generality of the proposed SHOE-CNN. As our proposed SHOE-CNN takes the semantically-close visual attribute into consideration, it is specially applied to the multi-class attribute. Three multi-class attribute types from [39] are used for evaluation, including "Neckline", "Sleeves Length", and "Category". For the attribute type such as "Neckline", we consider "Round Neckline" and "V-Shape Neckline" as one SA and the rest as

Fig. 7. Example retrieval results with top-8 returned shoe candidates. The correctly retrieved results with exactly the same shoe to the query are highlighted in green.

TABLE III
COMPARISON OF THE MULTI-CLASS ATTRIBUTE PREDICTION ACCURACY ON CLOTHING ATTRIBUTE DATASET [39]

	Chen *et al.* [39]	CNN	SHOE-CNN
Neckline	53.2	54.2	**56.50**
Sleeves	67.8	69.0	**71.20**
Category	47.5	53.4	**55.20**

TABLE IV
SHOE DETECTION PERFORMANCE OF REAL-WORLD SHOE PHOTOS COMPARING WITH TWO BASELINES: CNN DETECTION AND DPM DETECTION

Method	IoU Score
DPM [12]	53.22
CNN Detection	58.31
Proposed Method	**73.15**

another SA. Table III demonstrates the comparison results, and we can find that the proposed SHOE-CNN improves consistently over the conventional CNN by 2.3% for each of the three multi-class attributes. Moreover, the attribute prediction accuracy by SHOE-CNN outperforms that using traditional hand-crafted features by 3.3%, 3.4%, and 7.7%.

3) Shoe Region Proposal Selection: Based on the confidence scores from the three models and the learnt weights, a list of ranked region proposals are generated. The detection performance is calculated in terms of the IoU score between the selected region proposals and the ground truth bounding box. The top 3 region proposals for each shoe image are kept for the following experiment. We use the DPM detection and CNN detection as the baseline comparison to verify the effectiveness of our proposed region proposal selection scheme. From the results shown in Table. IV, it can be seen that our proposed method improves about 20% with respect to the original DPM detection and outperforms the CNN detection by about 15%.

4) Cross-Domain Shoe Retrieval: To quantitatively evaluate the effectiveness of the proposed SHOE-CNN and the discriminative multi-scale feature representation, we use several techniques to generate different features and compare their retrieval performances. The performance is evaluated in terms of the top-K retrieval accuracy. If the top-K returned results contain exactly the same shoe to the query item, then it is considered as successful.

We compare our proposed method (Level 1 + Level 2 + Level 3 feature (w/ SA) + PCA + Similarity Metric Learning) with both the traditional state-of-the-art pipelines and the deep learning based approaches. Two representative instance retrieval approaches with traditional features are used as baselines for comparison: 1) GIST feature [4] with 1024-dimension; 2) Dense SIFT feature followed by fisher vector encoding (DSIFT + Fisher Vector) [13] with the codebook size set as 64. The deep features extracted from the pre-trained CNN network on the whole image are used as a baseline to demonstrate the advantage of our proposed SHOE-CNN over the conventional CNN architecture. We also compare each level feature from SHOE-CNN (w/ SA) with that extracted from fine-tuned CNN without SA (w/o SA). The performance of multi-level deep feature representation is compared to that with the features from different levels individually. To further evaluate the effectiveness of our proposed system, we also compare it with a recent work on the

TABLE V
TOP-K RETRIEVAL ACCURACY OF OUR SYSTEM AND BASELINE RESULTS

Feature Configurations	Dim	Top-1	Top-5	Top-10	Top-20	Top-30
Gist feature [4]	1,024	4.12	7.03	8.31	10.03	11.41
DSIFT + Fisher Vector [13]	1,024	11.11	13.92	17.31	18.54	22.15
Deep Pre-trained CNN (WI)	4,096	11.87	19.22	23.94	28.28	31.59
Metric Network [49]	4,096	21.60	36.65	42.86	48.63	52.42
Level 1 feature (w/o SA)	5,376	16.14	25.29	29.83	35.53	38.92
Level 1 feature (w/ SA)	5,376	18.21	26.33	31.85	37.02	41.31
Level 2 feature (w/o SA)	5,376	15.37	23.30	27.17	32.92	36.61
Level 2 feature (w/ SA)	5,376	19.34	27.44	33.04	38.76	42.90
Level 3 feature (w/o SA)	1,024	15.23	22.76	26.84	31.68	34.69
Level 3 feature (w/ SA)	1,024	16.04	25.23	31.45	35.17	38.40
Level 1 + Level 2 + Level 3 feature (w/o SA)	11,776	22.37	37.58	43.70	49.49	52.98
Level 1 + Level 2 + Level 3 feature (w/ SA)	11,776	25.19	40.69	47.74	53.77	57.20
Level 1 + Level 2 + Level 3 feature (w/ SA) + PCA	1,024	26.47	44.41	50.67	57.62	60.94
Level 1 + Level 2 + Level 3 feature (w/ SA) + PCA + Similarity Metric Learning	1,024	**31.97**	**49.87**	**60.29**	**69.25**	**73.33**

cross-domain product retrieval [49]. Note that all the baselines except [49] utilize the cosine similarity as the metric.

Table V presents the results of our method and several baselines with different values of K. According to the experimental results, the deep features has a large improvement over the state-of-the-art traditional system utilizing DSIFT + Fisher Vector. The deep features (w/ SA) extracted from SHOE-CNN improve consistently over those without the subordinate attribute (w/o SA). Specially, the Level 2 feature (w/ SA) is better than Level 2 feature (w/o SA) by about 5.8% (top-20 accuracy), which demonstrates the effectiveness of the SHOE-CNN guided by the subordinate shoe attribute learning. Moreover, the Level 2 (w/ SA) feature achieves the best performance when compared to that with features from the other two levels. The three-level feature with SA has a large improvement over the best-performing single-level feature, Level 2 (w/ SA). By further integrating OASIS similarity metric learning, the accuracy further improves about 12% (top-20 accuracy) by using cosine similarity with the same feature. The top-30 retrieval accuracy of our proposed algorithm almost achieves 73%. This is useful for the real-world retrieval scenario where the users are more interested in the top-ranked results.

Finally, we show example retrieval results in Fig. 7. The top three rows indicate the successful retrieval results while the last two rows show the error cases. The results demonstrate that our system is capable of handling images with human skin like the second example, some challenging pose (*e.g.*, the first example), also the background clutter (*e.g.*, the newspaper beside the wedge shoes in the third row). From the failure examples, we notice they are caused by large variation of viewpoints (*e.g.*, only the back view of the shoe appear in the fourth row), and negative impact of skin. Also it is still difficult for our proposed feature representation to capture the subtle decoration details (*e.g.* the number of buckles in the query image in the fifth row). Based on extensive observations of our retrieval results, we find that in most cases we can achieve promising performance on shoes which only occupy a small area of the query image and cluttered background; however, we fail at the query images with large area of human skin exposed.

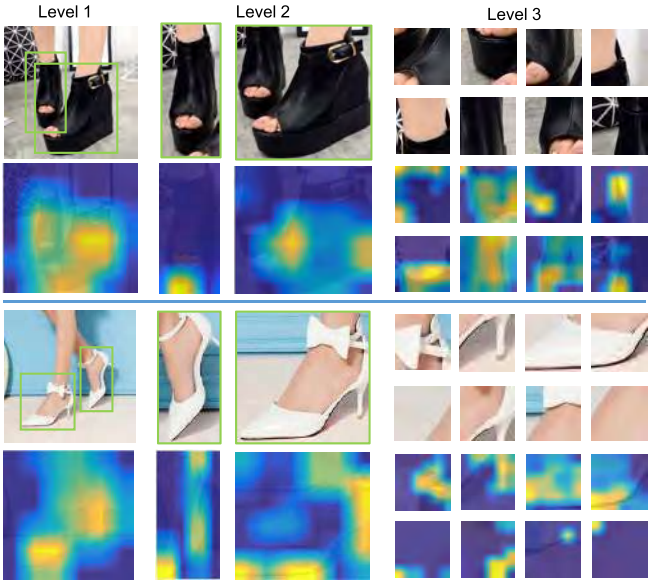
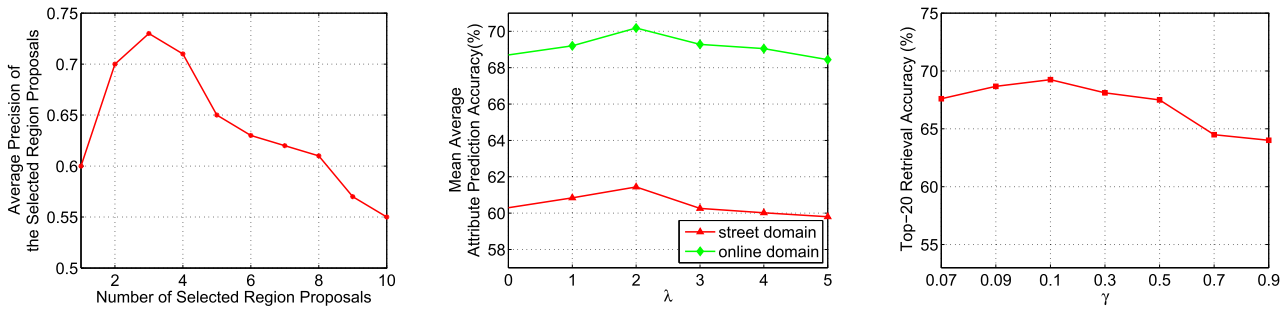


Fig. 8. SHOE-CNN feature maps. Row 1 and 3: two examples of the daily shoe photos (wedge and high-heel shoes). Row 2 and 4: corresponding activated feature maps from the pool5 layer for the whole query image (Level 1), the top-scored region proposals (Level 2) referring to each single shoe of a pair, and the local fine-grained patches (Level 3).

To provide an intuitive understanding about what the SHOE-CNN learns, Fig. 8 illustrates the activated feature maps with different levels of query image as the input. It can be seen that the feature maps for the whole image (Level 1) show strong activations on the overall structure of the shoes thus conveying the global information. For the feature maps with the selected high-scored region proposals (Level 2) and local parts (Level 3), generally they have strong activation values on the semantic parts of the shoes that convey the local appearances of the shoes and distinctive features (such as heel



(a) (b) (c)

Fig. 9. (a) Detection performance with different numbers of selected region proposals. (b) Mean average attribute prediction accuracy with respect to different values of λ. (c) Top-20 retrieval accuracy when we vary the parameter γ.

in the bottom row) for a unique type of shoes. Such learnt multi-level feature representation benefits the SHOE-CNN in describing both the global and local details of shoes.

5) Parameter Discussion: To further figure out the influence of parameter settings to the performance of our system, we evaluate the impact of the number of the selected region proposals on the detection accuracy, different settings of λ (Eq. 4) on the mean average attribute prediction accuracy, and γ (Eq. 13) on the top-20 retrieval accuracy, respectively, as shown in Fig. 9. Note that $\lambda = 0$ indicates the subordinate attribute is not considered. Fig. 9(a) demonstrates that the top-3 scored region proposals have the best detection performance compared to other values of the selected region proposals. It can be seen from Fig. 9(b) that on average incorporating the subordinate attribute improves attribute prediction accuracy. We notice that $\lambda = 2$ achieves the best results. However, as λ further increases, the prediction accuracy begins to drop. It indicates that the common semantic attribute prediction is also helpful to achieve good attribute prediction results and we should balance that with the subordinate attribute prediction. As shown in Fig. 9(c), the metric learning significantly boosts the retrieval performance (as shown in Table V) and the best performance is achieved with $\gamma = 0.1$.

6) Computational Complexity: The experiments of the proposed cross-domain shoe retrieval system are performed on a PC with an Intel Xeon E5-2630 CPU with 6 cores, 96G RAM, and a Tesla K40 GPU. For the shoe region proposal selection part, the training for the CNN detection model takes about 3.5 hours and the DPM model costs about 2.2 hours; for the shoe retrieval part, the training for SHOE-CNN takes about 5.8 hours and 4 hours for the online domain and the street domain, respectively. Given a query image, the proposed region proposal selection algorithm returns the top-3 scored high-quality region proposals in about 0.73s (Edgebox: 0.15s, DPM: 0.48s, CNN detection model: 0.1s); given the top-3 scored region proposals, it costs about 0.26s for feature extraction and shoe retrieval in our experiment.

VI. Conclusion

We present a cross-domain shoe retrieval system which aims at finding the exact same shoe image in online shops given a daily life query image. The SHOE-CNN network with a newly designed loss function by investigating the hierarchical properties of shoe attributes is proposed to extract a three-level feature representation for shoes. We show that our proposed three-level feature representation with the SHOE-CNN activated deep feature outperforms other feature configurations. Moreover, we also build a large-scale shoe image dataset consisting of 14341 street domain and 12652 online domain shoe images. Experiments carried on our dataset demonstrate the effectiveness of our proposed retrieval system. Comparing features extracted from pre-trained AlexNet, the accuracy of our system improves over 40%.

There are still several directions for further improvements for which we consider as our future work: 1) To be more consistent with the real-world applications, we will expand our dataset by crawling more images in our reference image set to include shoe photos from different views; 2) The performance of our retrieval system degrades when the human skin is exposed, especially when the feet is partially wrapped by shoes, such as the query shoe image in Fig. 3. Thus, we need to develop a human skin detection and exclusion technique to mitigate the negative effect of human skin without losing the details of shoes.

Acknowledgment

This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore.

References

[1] *Global eCommerce Sales, Trends and Statistics 2015*. Accessed on Sep. 2, 2015. [Online]. Available: https://www.remarkety.com/global-ecommerce-sales-trends-and-statistics-2015

[2] (2017). *The E-Boom of Online Shoe Sales is Taking Off*. Accessed on May 17, 2017. [Online]. Available: http://iagori.com/the-e-boom-of-online-shoe-sales-is-taking-off/

[3] L. Xie, J. Wang, B. Zhang, and Q. Tian, "Fine-grained image search," *IEEE Trans. Multimedia*, vol. 17, no. 5, pp. 636–647, May 2015.

[4] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[5] A. K. Jain and A. Vailaya, "Image retrieval using color and shape," *Pattern Recognit.*, vol. 29, no. 8, pp. 1233–1244, Aug. 1996.

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[7] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[8] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[9] A. Joly and O. Buisson, "Logo retrieval with a contrario visual query expansion," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 581–584.

[10] A. Yu and K. Grauman, "Fine-grained visual comparisons with local learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 192–199.

[11] T. Berg, J. Liu, S. W. Lee, M. L. Alexander, D. W. Jacobs, and P. N. Belhumeur, "Birdsnap: Large-scale fine-grained visual categorization of birds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2011–2018.

[12] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[13] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1704–1716, Sep. 2012.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[16] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. (2014). "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs." [Online]. Available: https://arxiv.org/abs/1606.00915

[17] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[18] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 584–599.

[19] A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson. (2014). "A baseline for visual instance retrieval with deep convolutional networks." [Online]. Available: https://arxiv.org/abs/1412.6574

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 346–361.

[21] L. Zheng, Y. Zhao, S. Wang, J. Wang, and Q. Tian. (2016). "Good practice in CNN feature transfer." [Online]. Available: https://arxiv.org/abs/1604.00133

[22] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 392–407.

[23] H. Yao, S. Zhang, Y. Zhang, J. Li, and Q. Tian, "Coarse-to-fine description for fine-grained visual categorization," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4858–4872, Oct. 2016.

[24] J. Mei, M. Liu, H. R. Karimi, and H. Gao, "Logdet divergence-based metric learning with triplet constraints and its applications," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4920–4931, Nov. 2014.

[25] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, "Large scale online learning of image similarity through ranking," *J. Mach. Learn. Res.*, vol. 11, pp. 1109–1135, Jan. 2010.

[26] J. Wang *et al.*, "Learning fine-grained image similarity with deep ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1386–1393.

[27] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang, "Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4766–4779, Dec. 2015.

[28] P. Wu, S. C. H. Hoi, H. Xia, P. Zhao, D. Wang, and C. Miao, "Online multimodal deep similarity learning with application to image retrieval," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 153–162.

[29] K. Sohn, "Improved deep metric learning with multi-class N-pair loss objective," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1849–1857.

[30] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1778–1785.

[31] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 951–958.

[32] T. Berg and P. N. Belhumeur, "POOF: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 955–962.

[33] A. Angelova and S. Zhu, "Efficient object detection and segmentation for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 811–818.

[34] A. Vedaldi *et al.*, "Understanding objects in detail with fine-grained attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3622–3629.

[35] K. Li, K. Pang, Y.-Z. Song, T. Hospedales, H. Zhang, and Y. Hu, "Fine-grained sketch-based image retrieval: The role of part-aware attributes," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–9.

[36] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan, "Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3330–3337.

[37] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, "PANDA: Pose aligned networks for deep attribute modeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1637–1644.

[38] J. Huang, R. S. Feris, Q. Chen, and S. Yan, "Cross-domain image retrieval with a dual attribute-aware ranking network," in *Proc. Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1062–1070.

[39] H. Chen, A. Gallagher, and B. Girod, "Describing clothing by semantic attributes," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 609–623.

[40] Y. Wang and G. Mori, "A discriminative latent model of object classes and attributes," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 155–168.

[41] K. Yamaguchi, M. H. Kiapour, and T. L. Berg, "Paper doll parsing: Retrieving similar styles to parse clothing items," in *Proc. Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3519–3526.

[42] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3570–3577.

[43] Z. Zhou, J. Zhou, and L. Zhang, "Demand-adaptive clothing image retrieval using hybrid topic model," in *Proc. ACM Multimedia Conf.*, 2016, pp. 496–500.

[44] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1096–1104.

[45] Y. Wang, S. Li, and A. C. Kot, "On branded handbag recognition," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1869–1881, Sep. 2016.

[46] Q. Chen, J. Huang, R. Feris, L. M. Brown, J. Dong, and S. Yan, "Deep domain adaptation for describing people based on fine-grained clothing attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5315–5324.

[47] A. Kovashka, D. Parikh, and K. Grauman, "WhittleSearch: Image search with relative attribute feedback," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2973–2980.

[48] J. Huang, S. Liu, J. Xing, T. Mei, and S. Yan, "Circle & search: Attribute-aware shoe retrieval," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 11, no. 1, Aug. 2014, Art. no. 3.

[49] M. H. Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg, "Where to buy it: Matching street clothing photos in online shops," in *Proc. Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3343–3351.

[50] A. Vedaldi and K. Lenc, "MatConvNet: Convolutional neural networks for MATLAB," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 689–692.

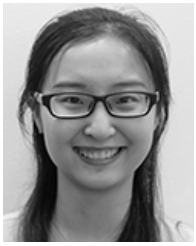
[51] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 391–405.

[52] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: From pairwise approach to listwise approach," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 129–136.

[53] X. Gao, S. C. Hoi, Y. Zhang, J. Wan, and J. Li, "SOML: Sparse online metric learning with application to image retrieval," in *Proc. Assoc. Adv. Artif. Intell.*, 2014, pp. 1206–1212.

[54] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," *J. Mach. Learn. Res.*, vol. 7, pp. 551–585, Dec. 2006.

[55] M. Ozeki and T. Okatani, "Understanding convolutional neural networks in terms of category-level attributes," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 362–375.



Huijing Zhan (S'17) received the B.S. degree from the Special Class for the Gifted Young, Huazhong University of Science and Technology, Wuhan, China, in 2012. She is currently pursuing the Ph.D. degree in electrical and electronic engineering with Nanyang Technological University, Singapore.

She was an Exchange Student with the AOTULE Graduate Student Research Exchange Program, Tokyo Institute of Technology, Tokyo, Japan, in 2013. Her research interests include computer vision, product retrieval, and fine-grained object recognition.

Boxin Shi (M'14) received the B.E. degree from the Beijing University of Posts and Telecommunications in 2007, the M.E. degree from Peking University in 2010, and the Ph.D. degree from The University of Tokyo in 2013. After his post-doctoral research at MIT Media Lab, Singapore University of Technology and Design, and Nanyang Technological University from 2013 to 2016, he joined the National Institute of Advanced Industrial Science and Technology as a Researcher. His research interests include computational photography and computer vision. He was a recipient of the Best Paper Runner-Up Award at the International Conference on Computational Photography 2015.

Alex C. Kot (S'85–M'89–SM'98–F'06) has been with Nanyang Technological University, Singapore, since 1991. He was the Head of the Division of Information Engineering with the School of Electrical and Electronic Engineering for eight years, and also served as the Associate Chair/Research and the Vice Dean Research for the School of Electrical and Electronic Engineering. He is currently a Professor and the Associate Dean of the College of Engineering and the Director of the Rapid-Rich Object Search Laboratory. His research interests include signal processing for communication, biometrics, data-hiding, image forensics, information security, and image object retrieval and recognition.

He is a fellow of IES and a fellow of the Academy of Engineering, Singapore. He is the IEEE Distinguished Lecturer of the Signal Processing Society. He served as an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE SIGNAL PROCESSING LETTERS, the IEEE SIGNAL PROCESSING MAGAZINE, the IEEE JOURNAL OF SPECIAL TOPICS IN SIGNAL PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS I: FUNDAMENTAL THEORY AND APPLICATIONS, and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II: ANALOG AND DIGITAL SIGNAL PROCESSING. He has served the IEEE Signal Processing Society in various capacities, such as the General Co-Chair at the 2004 IEEE International Conference on Image Processing and the Vice President of the IEEE Signal Processing Society. He was a recipient of the Best Teacher of the Year Award and co-authored several best paper awards, including for ICPR, IEEE WIFS, and IWDW.