

# TOWARDS ATTRIBUTE-AWARE CROSS-DOMAIN IMAGE RETRIEVAL

**JUNSHI HUANG**

*(M.Sc, Major in Computer Science, Beijing Institute of Technology)*

A THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF  
PHILOSOPHY



Department of Electrical and Computer Engineering


National University of Singapore

2015

# Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

A handwritten signature in black ink, appearing to read 'Huang Junshi'.

Junshi Huang

Aug. 2015

# Acknowledgement

I would like to appreciate those people who have helped and supported me during my PhD study.

Foremost, I want to express my special gratitude to my supervisor Dr. Shuicheng Yan, who accepted and supported me as his student without any hesitation after our first discussion. I would like to thank him for encouraging and instructing my research works. His tremendous knowledge, constructive suggestions greatly helped me to lay down the foundation on my research. I have been deeply inspired by his dedication to research, his patient and enthusiastic on instructing students, which keep benefiting me for my whole life.

Meanwhile, I want to thank my former supervisor Dr. Ying Chen in BIT and mentor Dr. Jun Yan in MSRA for providing me with the opportunity and courage to start my PhD education. I also want to give my thanks to Dr. Rogerio Feris in IBM for his kind help on my research work.

Then, I would like to thank our lab members for their great help in my four years study and life in NUS. Specially, I want to thank Dr. Qiang Chen, Dr. Junliang Xing, and Dr. Si Liu for their patient and helpful guidance. Particularly, I am very grateful for their always support in my life. I also want to thank my good friends Ming Chen, Csaba Domokos, Jian Dong, Min Lin, Luoqi Liu, Wei Xia, and Zhao Zhang (As I cannot tell who is more important to me, I list them in the ascending order of their last name). I still remember the sleepless nights we have spent together. Besides, I have met a lot of great friends in LV lab. Here, I want to express my sincere gratitude to all of you.

Last but not least, I want to thank my grandparents, my parents and my girlfriend for their everlasting support and care. Particularly, I want to dedicate this thesis to my grandfather. We miss you so much.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Thesis Focus and Main Contributions . . . . .	4
1.2	Organization of Thesis . . . . .	8
<b>2</b>	<b>Literature Review</b>	<b>9</b>
2.1	Retrieval Dataset Collection . . . . .	9
2.2	Object Detection . . . . .	10
2.3	Visual Image Features . . . . .	11
2.4	Describing Object by Attributes . . . . .	12
2.5	Deep Learning for Image Retrieval . . . . .	12
2.6	Domain Adaptation . . . . .	13
<b>3</b>	<b>Semantic Parts for Object Retrieval</b>	<b>15</b>
3.1	Introduction . . . . .	16
3.2	Related Work . . . . .	20
3.3	The Footwear Dataset . . . . .	21
3.3.1	Automatic Images Collection . . . . .	22
3.3.2	Fine-grained Attributes . . . . .	22
3.3.3	Semantic Footwear Parts . . . . .	23
3.4	The Circle & Search System . . . . .	24
3.4.1	Attribute-aware Detection . . . . .	25
3.4.2	Query-Specific Attribute Refinement . . . . .	34
3.4.3	Extension of Retrieval System . . . . .	35
3.5	Experiments . . . . .	36
3.5.1	Exp1: Semantic Footwear Detection . . . . .	36



3.5.2	Exp2: Attribute-aware Footwear Retrieval . . . . .	41
3.6	Chapter Summary . . . . .	46
<b>4</b>	<b>Describing People by Clothing Attributes</b>	<b>47</b>
4.1	Introduction . . . . .	48
4.2	Related Work . . . . .	51
4.3	Dataset Preparation . . . . .	53
4.3.1	Online Shop Dataset . . . . .	53
4.3.2	Offline Street Dataset . . . . .	54
4.4	Approach . . . . .	55
4.4.1	RCNN for Body Detection . . . . .	56
4.4.2	Deep Domain Adaptation . . . . .	57
4.5	Experiments . . . . .	60
4.5.1	Implementation Details . . . . .	60
4.5.2	Exp1: RCNN body detection . . . . .	61
4.5.3	Exp2: DDAN for Fine-grained Attribute Classification . .	62
4.5.4	Application 1: Attribute-based People Search . . . . .	64
4.5.5	Application 2: Street2Shop Clothing Retrieval . . . . .	65
4.6	Chapter Summary . . . . .	65
<b>5</b>	<b>Cross-domain Attribute-aware Image Retrieval</b>	<b>67</b>
5.1	Introduction . . . . .	68
5.2	Related Work . . . . .	71
5.3	Data Collection . . . . .	74
5.4	Approach . . . . .	75
5.4.1	Dual Attribute-aware Ranking Network . . . . .	76
5.4.2	Clothing Detection . . . . .	80
5.4.3	Cross-domain Clothing Retrieval . . . . .	80
5.5	Experiments . . . . .	82
5.5.1	Experimental Setting . . . . .	82
5.5.2	Clothing Detection Improving Clothing Retrieval Performance . . . . .	83
5.5.3	Cross-domain Clothing Retrieval Evaluation . . . . .	84

5.5.4	Attribute-aware Clothing Retrieval Evaluation . . . . .	86
5.5.5	Showing the Robustness: Performance vs. Retrieval Gallery Size . . . . .	86
5.5.6	System Running Time . . . . .	87
5.6	Chapter Summary . . . . .	87
<b>6</b>	<b>Conclusions and Future Works</b>	<b>88</b>
6.1	Thesis Conclusions . . . . .	88
6.2	Future Works . . . . .	91

# Summary

Cross-domain image retrieval is a foundational yet important task for many practical applications, e.g., mobile product retrieval in online-shopping system, people re-identification, etc. In this thesis, we address the problem of cross-domain image retrieval based on the fine-grained attributes, and introduce the industrial applications of our framework on the web-scale cross-domain object retrieval system.

The cross-domain image retrieval is a challenging problem which is partial due to the large domain discrepancies in terms of changeable background, pose, and illumination conditions. We approach this problem by first collecting large-scale images depicting the online-offline object pairs, e.g., footwear, garment, and their fine-grained attributes. To eliminate the impact of cluttered background, the object detection systems are harnessed to crop the foreground objects. The traditional detection models, e.g., Deformable Part Model, and some emerging detection models, e.g., R-CNN and its variants, are introduced and enhanced in this component. Based on the aligned object patches, the framework for retrieval feature learning is jointly driven by the semantic attributes and visual appearance.

In the conventional retrieval framework, we characterize our retrieval system by using the attributes as the consistencies among several deformable parts, which are considered as the high-order semantic relations among parts. By maintaining these high-order relations, the inference of parts position and attributes can be mutually enhanced. After embedding the inferred attributes into the metric learning framework, the retrieval system ranks the images according to the refined attributes by taking advantage of attributes correlation. Most recently, inspired by the success of Convolutional Neural Network (CNN), we propose the

dual-structure attribute-aware network for the problem of cross-domain people description and clothes retrieval based on the fine-grained attributes. Basically, the dual-structure network is composed of two sub-networks, each of which models the domain-specific object by feeding images from one domain. The alignment layers in-between two sub-networks constrain the comparability of responses of two sub-networks, and thus guarantee the domain adaptation of features. By using the attributes as alignment constraint or semantic embedding, the semantic representation of object is precisely learned as well. Referring to different tasks, we design task-specific loss functions and alignment layers, and achieve state-of-the-art performance on the people description and clothes retrieval problem in our experiments.

In summary, we propose several novel frameworks for the attribute-aware cross-domain image retrieval in this thesis. The experiments demonstrate the promising performance of our proposed methods, as well as their potential in industrial applications.

# List of Tables

3.1	The comparison of detection performance between baseline and our method in different settings. . . . .	38
3.2	Attribute classification accuracy of three baselines and our method.	40
3.3	The average score and standard deviation of user study on the baselines and our retrieval system . . . . .	44
4.1	Fine-grained attributes classification results for <i>Street-data-a</i> . . .	62
4.2	Fine-grained attributes classification results for <i>Street-data-c</i> . . .	63
5.1	Clothing attribute categories and example values. The number in brackets is the total number of values for each category. . . . .	73
5.2	AP of detection models on online-offline images and its corresponding top-20 retrieval accuracy on a subset of the data. . . .	83
5.3	The NDCG@20 result evaluating the attribute level match on 200,000 retrieval gallery. . . . .	87

# List of Figures

1.1	A general framework of image retrieval . . . . .	2
1.2	Overview of cross-domain image retrieval. The discrepancies of online and offline images are bridged by the fine-grained attributes. The cross-domain image retrieval is defined as follows: given the query image in one domain, e.g., a street photo or surveillance frame, retrieve the visually or semantically similar online images in another domain, e.g., product images in an online shop dataset.	4
3.1	Scenario illustration of the Circle & Search footwear retrieval system. The user browses a website, and circles a shoe. The visually and semantically similar footwear images will be returned by the proposed system from the retrieval repository. . . . .	16
3.2	The semantic parts of footwear in three viewpoints are presented in (a). The color of each part indicates its corresponding attribute. The mapping of part color and attributes can be seen in the right part of (b). The semantic meaning of those parts is introduced in Sec 3.3.3. The number within each part indicates its order in the tree-structure model. By applying different combination strategies of manually labelled part in the detection model, we present two well-designed tree-structures for profile-view and frontal-view, respectively. Similarly, the color of each part indicates its corresponding attributes. . . . .	18

3.3	Some footwear examples of different viewpoints in our dataset. The images in the first row are the product images; The images in the second row are daily photos. Totally, our dataset contains 17,151 footwear images of several viewpoints. . . . .	22
3.4	The illustration of footwear attributes. The first column of each table is the attributes category, and the rest columns of each table are the corresponding attribute values. . . . .	23
3.5	The framework of our proposed footwear retrieval system: given a footwear image (a), the parts (c) and attributes (d) of the footwear instance are predicted by our attribute-aware part-based detection model (b) in an iterative manner. These attributes are then fed into a query-specific attribute refinement retrieval model (e) for refinement. (f) The refined attributes are used as retrieval feature to retrieve the retrieval results. . . . .	25
3.6	The sub-figure (a) presents the “AND-OR” hierarchical structure of 10-th part. Firstly, the 10-th part of training samples are di- vided into several components with “AND” relation defined by at- tribute values, which constructs the first-level mixture (a). Each component of first-level mixture is consisted of $K$ components with “OR” relation generated by $K$ -means (b). The cross at the origin of coordinates in sub-figure (b) represents the position of 10-th part’s parent node, <i>i.e.</i> , 9-th part, and the features of parts for $K$ -means are the normalized distance along the x-axis and y-axis from 10-th part to 9-th part. . . . .	26
3.7	Some examples of the detected bounding boxes. In the result, we can observe that our detection model can effectively localize the different footwear parts even the scale, viewpoint are quite diverse or background is cluttered. . . . .	40

3.8	The nDCG@k of baselines and our proposed retrieval system, <i>i.e.</i> , query-specific attribute refinement retrieval method. The experiment results present that nDCG@k of specific retrieval method gradually decreases as the number of retrieval images increases, which matches our expectation as the later retrieval images usually are more irrelevant to query image. When fixing the number of retrieval images, our retrieval method outperforms the baselines under most of the experiment configurations. Specifically, two aspects can be observed. If we use the same input data as shown in each sub-figure, our query-specific attribute refinement method is superior than the other two searching strategies. If we use the same retrieval strategy, <i>i.e.</i> , the lines of the same color across sub-figure (a) to sub-figure (c), the retrieval method with our detected result (sub-figure (c)) outperforms the methods combined with detection baselines (sub-figure (a) and sub-figure (b)). This observation implicitly indicates the better performance of our detection method. In sub-figure (d), we can observe that the our retrieval result is also comparable to the state-of-the-art retrieval systems. . . . .	43
4.1	Overview of the proposed approach. We propose a novel deep domain adaptation method to bridge the gap between images crawled from online shopping stores and unconstrained photos. Another unique aspect of our work is the ability to describe people based on their fine-grained clothing attributes. . . . .	48
4.2	Overview of our proposed approach. . . . .	55
4.3	Enhanced R-CNN detection pipeline. . . . .	57
4.4	Precision-recall curves for body detection results on <i>Street-data-a</i> .	62



4.5	Application 1: Attribute-based people search. We rank the images according to their attribute scores. The top-5 ranking results for each query are exhibited. Top 2 rows results are from <i>Street-data-a</i> , and the bottom results are from <i>Street-data-c</i> . The images that exactly match the query are marked with red bounding box. Best viewed in original pdf file. . . . .	64
4.6	Application 2: Street2Shop clothing retrieval. Top 2 rows results are from <i>Street-data-a</i> , and the bottom two rows are from <i>Street-data-c</i> . We output the top 3 retrieval results for both datasets. Best viewed in original pdf file. . . . .	64
5.1	Cross-domain clothing retrieval. (a) Query image from daily photos. (b) Top-8 product retrieval results from the online shop domain. The proposed system finds the exact match clothing (first two images) and ranks the ones with similar attributes as top results.	68
5.2	Some examples of online-offline image pairs, containing images of different human pose, illumination, and varying background. Particularly, the offline images contain many selfies with high occlusion.	73
5.3	The distribution of online-offline image pairs. . . . .	74
5.4	The specific structure of DARN, which consists of two sub-networks for images of the shopping scenario and street scenario, respectively. In each sub-network, it contains a NIN network, including all the convolutional layers, followed by two fully connected layers. The tree-structure layers are put on top of each network for attribute learning. The output features of each sub-network, i.e., FC1, Conv4-5, are concatenated and fed into the triplet ranking loss across the two sub-networks. . . . .	76
5.5	The top-k retrieval accuracy on 200,000 retrieval gallery. The number in the parentheses is the top-20 retrieval accuracy. . . . .	84
5.6	The top-20 retrieval accuracy on different sizes of retrieval galleries.	85

5.7	The top-4 retrieval result of DARN with Conv4-5. The images in first column are the queries, and the retrieved images with green boundary are the same clothing images. Best viewed in original pdf file. . . . .	86
-----	---	----

# Chapter 1

## Introduction

The image retrieval systems help people to search, browse, and make use of the digital images from the large-scale image database. The studies of image retrieval, involving feature learning, indexing, and matching (see Figure 1.1), have been very active since the 1970's. Generally, the methods for image retrieval can be categorized into the text-based and the content-based image retrieval. A popular pipeline for text-based image retrieval is to annotate the images by meta-data, e.g., tag, caption, or description of images, so that the retrieval problem can be solved over the annotation text. In the early 1990's, the content-based image retrieval was proposed due to the intrinsic drawbacks of text-based retrieval systems on the increasing number of image collections, e.g., unaffordable annotation cost, imprecise annotation labels, etc. Instead of text-based keywords, the content-based image retrieval methods extract the retrieval features from the visual content of images, including color, texture, and shape. In this thesis, we focus on the problem of cross-domain image retrieval, which is defined as follows: given a query image of domain  $X$ , retrieve similar images from the retrieval repository of domain  $Y$ . Particularly, we are interested in retrieval feature learning based on the image content driven by semantic attributes. Meanwhile, we apply our retrieval system on some industrial applications by taking the concrete objects as examples, and demonstrate the effectiveness of our system on the general object retrieval.

The cross-domain image retrieval is one of the most important yet challenging problems in the image retrieval domain. Many practical applications can

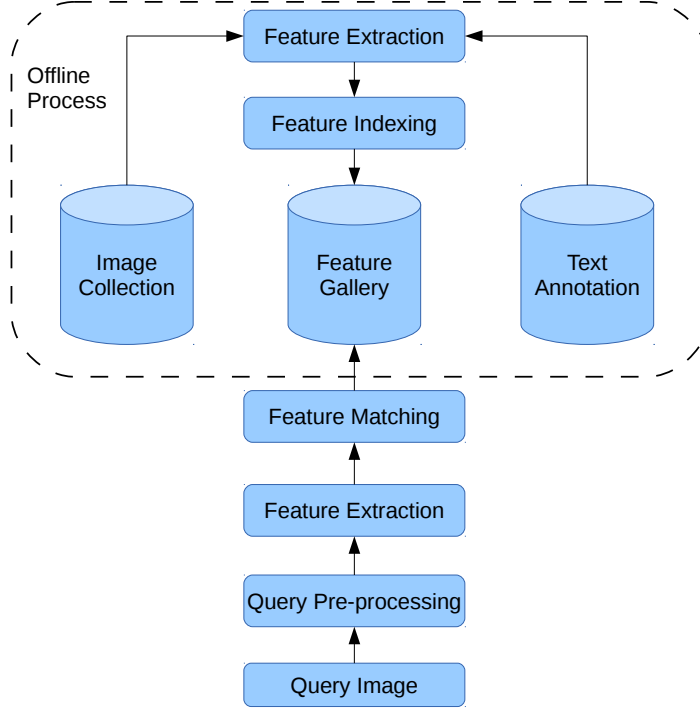


Figure 1.1: A general framework of image retrieval

benefit from the methods of cross-domain image retrieval. For instance, the customers browsing online-shopping websites aim to buy desired products depicting in a street photo. In the surveillance domain, the security guard may be interested in retrieving a suspect from cameras, given the query image from another camera. In a photograph exhibition, the visitors want to search the photos of a specific architecture taken by different photographers. In those scenarios, a distinct characteristic of the image retrieval problem is that the query image and the retrieval image come from different domains. Therefore, some existing obstacles resulted by the domain discrepancy must be removed before applying the retrieval techniques on practical applications. Usually, the images in different scenarios have large discrepancies in terms of changeable illumination, different resolutions, variant viewpoints, and cluttered background, which result in the incomparability of hand-craft image features. Though some domain adaptation methods [11, 51] can be used to bridge the gap between domains, they are still not effective enough to be directly applied on the industrial retrieval systems. Furthermore, the existing hand-craft features, e.g., SIFT, HOG, and LBP etc., capturing the visual appearance information of images are feeble in handling the

retrieval problem on non-rigid objects.

To address the aforementioned problems, the object detection [32, 39] or segmentation [120, 117, 119, 118] can be harnessed to eliminate the impact of cluttered background by cropping the foreground objects and their surrounding context from the whole images. Compared with object segmentation, some popular detection methods, e.g., Deformable Part Model (DPM) [32], R-CNN [39], are usually introduced and enhanced for the object detection due to their relatively high efficiency. Based on the detected objects, the attribute information can be embedded into the learning of retrieval features for the representation of objects at the semantic level, which may somehow attenuate the inconsistency of visual appearance caused by domain discrepancy. In this way, both the visual content and semantic information of objects can be captured by the learned features simultaneously. Furthermore, the domain adaptation methods can be applied on the top of learned features to further refine the retrieval results.

Inspired the success of semantic attributes in many applications [35, 70, 28, 68, 92, 8, 110, 53, 9], we aim to integrate the semantic attributes with visual appearance for image retrieval. Firstly, we suggest that the attributes can be attached into several parts of object, where the attributes and visual parts achieve a consistent relations. In other words, the semantic meaning of attributes and visual appearance of parts should agree with each other. Obviously, this constraint can effect the localization of object parts and prediction of attributes until the inference reaches a stable stage, where we claim that these alternative procedures can enhance each other. Recently, the CNN-based methods have achieved excellent performance on many computer vision problems [65, 39, 38, 73, 85, 124], and rare of them are proposed for image retrieval. We study this problem by integrating the attribute learning method into CNN framework for image retrieval problem. In our second work, we propose to use attributes as similarity criterion for the learning of retrieval features. Generally, the dual-path network structure is proposed for the learning of domain-specific images. At the high-level component of network, an alignment layer is added for the domain adaptation, where the attributes is embedded in a supervised mannar. Note that besides the supervised mannar, the similarity of images can also be measured by the

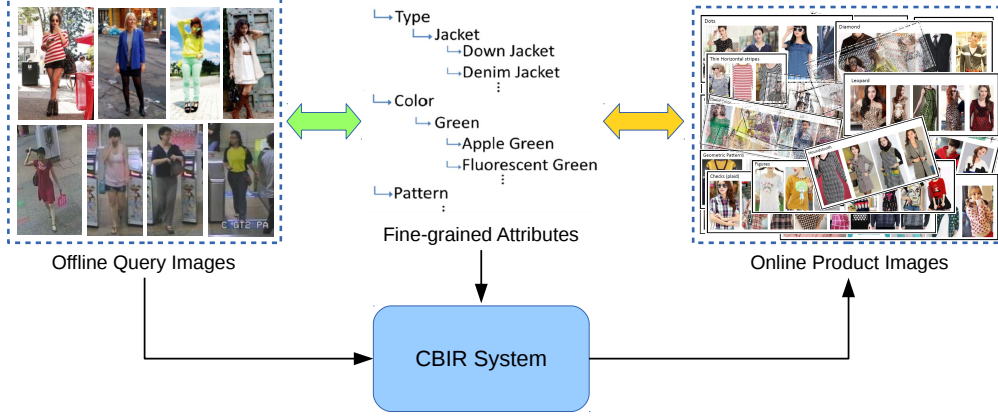


Figure 1.2: Overview of cross-domain image retrieval. The discrepancies of online and offline images are bridged by the fine-grained attributes. The cross-domain image retrieval is defined as follows: given the query image in one domain, e.g., a street photo or surveillance frame, retrieve the visually or semantically similar online images in another domain, e.g., product images in an online shop dataset.

output of network in an iterative unsupervised way. Analogously, we design another dual-structure network for the clothing images retrieval in our third work. Different from the second one, this structure extracts the semantic representation of images by learning the individual information and relations of multiple attributes in a holistic model. Based on the high-level representation of images, the triplet images are fed into the network for learning to rank, which is more intuitive in real applications.

This thesis studies the problem of cross-domain object retrieval by learning the attribute-aware retrieval features (see Figure 1.2). Meanwhile, we collect and release several large-scale fashion-related databases with fine-grained attributes for our retrieval systems and further academic research. The works introduced in this thesis are operated during 2013-2015.

## 1.1 Thesis Focus and Main Contributions

The standard retrieval system follows the pipeline of interested object detection, retrieval feature learning, feature indexing, and feature matching/ hashing. In this thesis, we focus on the object detection and retrieval feature learning motivated by the following facts:

1. In object detection, some popular approaches, e.g., DPM, and R-CNN,

have attracted lots of attentions in the object recognition community. However, the performance of those detection models are greatly limited by the image features, which only capture the appearance information of objects. By describing the objects with attributes and context information, we can learn a more discriminative feature at the semantic level for object detection.

2. Though lots of efforts have been made for the problem of large-scale image retrieval, few of them focus on the cross-domain image retrieval, which is more relevant to the practical applications. Some methods directly apply the domain adaptation methods on existing retrieval features, yet fail to utilize the correlation information between domains, e.g., object attributes, online-offline image pairs.
3. In attribute analysis, previous works only deal with the small set of coarse attributes describing the global properties of objects, e.g., red skirt, stripe shirt, which is insufficient to present the object details. To describe the object specifically, the fine-grained attributes and local information can be applied for the object representation learning.

To address the aforementioned drawbacks, this thesis introduces the concept of attribute-aware image representation for both object detection and image retrieval. Based on the semantic representation of objects, we propose to learn the retrieval feature in the source domain and target domain simultaneously. The main contributions of our works are listed as follows:

1. **Cross-domain Dataset and Fine-grained Attributes.** We collect a series of unique datasets composed of cross-domain image pairs with fine-grained attributes. The objects in the datasets involve the fashion-related objects, e.g., clothes, footwear, and pedestrian in the surveillance videos. Overall, there are millions of online images and tens of thousands of offline counterparts in the datasets. Each image has about 5-10 semantic attribute categories, with more than a hundred attribute values. After pruning the noisy labels, we qualitatively analyse the semantics of attributes, and organize those attributes into a semantic taxonomy. Along with the taxonomy

of attributes, these datasets composed of fashion objects under variant conditions provide rich information for the real-world applications.

2. **Semantic Object Detection.** To apply the object detection methods on the domain unconstrained images, we propose the attribute-aware object detection methods exploiting the semantic representation of images for object detection. In the conventional pipeline, we train our detection model on the foundation of DPM by associating the attributes with some relevant parts. The attributes in this model are used as the high-level constraints among relevant parts, requiring that the appearance of parts should be consistent enough to represent the semantic meaning of attributes. By applying the Expectation-Maximization (EM) approach, the detection model converges into a stable state where the visual parts and semantic attributes are highly unified.

In the R-CNN framework, we carefully design the CNN model based on the NIN structure for the detection feature extraction. Instead of training the model on ImageNet for visual representation, we propose the tree-structure layers for multi-attribute learning. In this structure, the low-level representations of images for each attribute are shared and fed-forward until splitting into several branches in the high-level layers. Each branch is a sub-network modeling one attribute individually. During training, the gradients from every branch are merged together in the splitting layer, so that the response from this layer captures the semantic information of multi-attributes simultaneously. By using the response of the attribute-aware network, the performance of object detection is greatly enhanced in both online and offline domains.

3. **People Description with Fine-grained Attributes.** Describing people in details is an important task for many applications, e.g., identifying a target suspect or finding missing people in the surveillance videos. The fine-grained attributes usually provide useful information in people description and re-identification. To fully exploit the attributes, we propose a Double-path deep Domain Adaptation Network (DDAN) to learn the



information from two domains jointly. In this network, some alignment cost layers are introduced and embedded into the two columns to guarantee the consistency of two domain features by using the prior information, e.g., attribute similarity or visual similarity. To deploy the model after co-training, a merging layer is designed to take the input of each column, and outputs the maximal response as feature activation. Compared with traditional domain adaptation approaches, DDAN aims to learn the aligned features from source and target domains in high-level layers, which achieves superior performance than that of subspace based methods [42, 56].

4. **Cross-domain Image Retrieval.** Due to the large discrepancies of images in different domains, the Dual Attribute-aware Ranking Network (DARN) is proposed for retrieval feature learning. Similar to DDAN, DARN consists of two sub-networks, one for each domain. Rather than directly using attributes for feature alignment, this network explores the attributes by the tree-structure layers for semantic representation learning. Based on the learned semantic representation, a triplet similarity constraint is harnessed for the learning to rank across two sub-networks, which ensures the comparability of domain-specific features. To further capture the subtle details of objects, we concatenate the responses from high-level convolutional layers, representing the local spatial information, and the responses from fully-connected layers, representing the global information, as retrieval features, which double the retrieval accuracy compared with state-of-the-arts.

Each of these ideas serves as one component of our cross-domain image retrieval system towards industrial requirements. As there may exist other works for the general image retrieval, our works may have bias towards some specific applications. In this thesis, most of the information, including cross-domain image pairs and fine-grained attributes, is collected from the Internet, and the annotators are only hired for the data curation.

## 1.2 Organization of Thesis

In Chapter 2, we elaboratively present the detailed literature reviews of relevant topics, including data collection, object detection, attribute embedding, and retrieval feature learning. In Chapter 3, we propose the semantic deformable part model for object detection by attributes and deformable parts, based on which the retrieval features are learned. In Chapter 4, the DDAN model is introduced for cross-scenario people description. In this chapter, the concept of the alignment layer implemented by the prior similarity function is proposed for domain adaptation. By using different side information for the similarity function, this model can be applied on both supervised or semi-supervised problems. Instead of directly using attribute in the similarity function, we learn the high-level semantic representation of images in Chapter 5, based on which the learning to rank framework is embedded to learn the retrieval features. This attribute-driven learning network, called DARN, achieves significant improvement in the cross-domain clothing retrieval problem. Lastly, we conclude our works and introduce the future works on cross-domain image retrieval in Chapter 6.

## Chapter 2

# Literature Review

Image retrieval is one of the most important and popular topics in computer vision and multimedia communities. Lots of great works [1, 102, 80, 54, 15, 12] have demonstrated effective methods for image retrieval. This chapter presents a survey of literature for the tasks in image retrieval in fashion domain and its relevant topics, including the collection of retrieval dataset, object detection, attribute analysis, deep learning, and domain adaptation methods. After introducing the “off-the-shelf” methods for each topic, we briefly present our ideas on the end-to-end retrieval system focusing on the dataset collection, semantic object detection and the learning of cross-domain retrieval features.

### 2.1 Retrieval Dataset Collection

Recently, several datasets containing a wide variety of online product images collected from several websites have been carefully annotated with textual labels [61, 100, 121, 80, 22, 55, 82, 8]. Particularly, Kang et al. [61] collected a large-scale dataset of about 5 million product images capturing 1.2 million objects of multiple views for the object discovery system. A special dataset for mobile product image search [100] is composed of 10 categories of products, and about 43,000 catalogue images. For fashion-relevant image retrieval, Yamaguchi et al. [121] created a novel dataset for studying clothing parsing, consisting of 158,235 fashion photos with associated text annotations. Liu et al. [80] proposed a dataset with 8,293 upper body images and 8,343 lower body images, both of

which are labeled with 15 clothing attributes. However, most of the datasets are constructed for the problem of image retrieval in single domain, and most of them are not large enough for real-world applications. To address the problem of large-scale cross-domain image retrieval, we crawled millions of fashion images, containing clothes, pants, and footwear, from many online-shopping websites, with the surrounding text parsed as attributes. The offline photos uploaded by the users are also collected from the customer review pages and are used as the offline counterparts of online images for the study of cross-domain image retrieval.

## 2.2 Object Detection

The object detection [30, 32, 39] aims to locate the position and scale of a specific object in the images. Traditionally, a set of outstanding detection methods, *i.e.* the pictorial model [30], and its variants [31, 32] use the mixture of deformable parts for the flexible representation of object. In inference, the object parts are densely placed on the images to find the high-energy regions by sliding windows. Though the pictorial model is effective on the object detection, the detection performance is greatly limited by the representation capability of hand-crafted features. Thereafter, the R-CNN framework [39, 38] takes advantage of deep features to improve the performance of object detection. In this framework, the objectiveness generation methods, *e.g.*, selective search [107], are used to generate the object candidates from raw images. After extracting the deep features of object candidates, the object detection problem can be modelled in the classification framework. However, the performance of R-CNN framework greatly depends on the recall of proposal generation method. Most recently, the updated version of R-CNN, named faster R-CNN [99], achieved an outstanding performance compared with many other deep feature based detectors. This framework simultaneously learns the Region-of-Interests (RoI) and classifies the RoI into foreground or background, which greatly reduce the dependence on region proposals. The superior performance of the faster R-CNN framework further indicates the higher representation capability of deep feature than that of

traditional features.

## 2.3 Visual Image Features

The canonical framework for many computer vision problems, e.g., object recognition, object detection, object segmentation, image retrieval, and scene understanding, etc. shares multiple similar modules, e.g., feature representation. In the feature representation part, a traditional pipeline is usually composed of three steps. Firstly, the descriptor extraction methods extract the low-level visual representation on the dense grids or sparse interest points of raw images, e.g., Histogram of Gradients (HoG) [13], Scale-Invariant Feature Transform (SIFT) [83], and Local Binary Patterns (LBP) [89], etc. Then, the feature encoding methods encode the low-level representation according to the predefined or learned dictionary. The recent coding methods can be categorized into several schemes, e.g., Vector Quantization (VQ) based, Sparse Coding (SC) based, and Gaussian Mixture Models (GMM) based. Particularly, the locally-constrained linear coding (LLC) [113] and Fisher kernel [95] are proved to be effective for many object recognition tasks [26, 20, 21]. The Vector of Locally Aggregated Descriptors (VLAD) [58] achieved superior performance on the image retrieval problem [2]. Finally, the encoded feature vectors are pooled together for the image-level representation according to different side information, e.g., the spatial information used in Spatial Pyramid Matching (SPM) [122] and Pyramid Matching Kernel (PMK) [46].

Most recently, the visual image features can be directly learned and extracted from the raw images by the deep learning framework, e.g., Convolutional Neural Network [19, 59]. Basically, a conventional convolutional network consists of several layers for a specific function, e.g., a convolutional layer for dense response extraction, a neuron layer for response activation, a pooling layer for feature pooling, and a fully-connected layer for global feature embedding, each of which is similar to the major steps in the traditional pipeline of feature extraction. In our work, we carefully design the structure of networks and objective functions according to the task-specific requirements for the learning of attribute-aware

deep features in our cross-domain image retrieval problem.

## 2.4 Describing Object by Attributes

The research topics on semantic attributes have attracted significant attentions in the computer vision community for decades [35, 70, 28, 68, 92, 8, 110, 53, 9]. Usually, the attributes are considered as the semantic-level properties of objects or scenes that are shared across categories or domains. Referring to the studies of attributes, lots of efforts have been made for the zero-shot learning [70], image retrieval [102, 63, 53], fine-grained categorization [6, 28], scene understanding [94], and semantic object description [67]. Besides, some research works [10, 17] focus on the attribute mining from noisy web data, which greatly benefit the studies of attributes on large-scale datasets.

Related to our work, Kovashka et al. [63] implemented a retrieval system, called “WhittleSearch”, by using the concept of “relative attributes” [92] for user interaction. Siddiquie et al. [102] exploited the co-occurrence of attributes for image ranking by multi-attribute queries. Some papers [5, 8, 72, 9, 101] explored the problem of clothes description and people re-identification by attributes. Specifically, Bourdev et al. [5] proposed to use the concept of poselet for people attribute classification, which elaboratively describes the details of people attributes. Analogously, Chen et al. [8] introduced a semantic attributes learning method for clothing by taking advantages of human pose estimation. Moreover, the mutual dependencies between attributes are explored by a Conditional Random Field. Similarly, our work follows the direction of semantic attribute learning yet focuses on the fine-grained attributes, which are proved to be more effective for the problem of cross-domain retrieval feature learning.

## 2.5 Deep Learning for Image Retrieval

The Deep Convolutional Neural Networks have made significant contributions to many computer vision tasks, e.g., image classification [65], object detection [39, 38], human parsing [78, 73], and many other computer vision applications [85, 124]. Meanwhile, some ideas on the network structure [40, 43, 74] and

regularization terms [87, 49] have been proposed for the improvement of general networks. Particularly, Lin et al. [74] proposed the Network-in-Network (NIN) structure for the richer modeling capability of the network by reconstructing the network response in the channel dimension. Though lots of great works [7, 104, 48] have been proposed for higher recognition performance, they are still suffer from the low efficiency in real-world application. To achieve the trade-off between effectiveness and efficiency, we integrate the NIN structure into our detection and retrieval framework for the sophisticated object representation.

In image retrieval, some deep learning based approaches [3, 64, 112] for the content-based image retrieval problem defeat the previous methods based on traditional image representation. However, those methods are not designed to handle the problem of cross-domain image retrieval. In contrast, we study the problem of deep learning methods on cross-domain image retrieval, and explicitly use the attributes as the side information for the learning of semantic features for domain adaptation.

## 2.6 Domain Adaptation

Many methods [41, 42, 45, 11, 51] have been proposed for the problem of domain adaptation based on the existing features. In the deep learning community, the domain adaptation methods are usually implemented by network fine-tuning, which retrain the last few layers or fine-tunes the whole network in the target domain with a smaller learning rate [41, 42, 45]. However, this method requires lots of training images, and the network performance in the source domain usually deteriorates after fine-tuning. A distinct method of domain adaptation is the Deep Learning for Domain Adaptation by Interpolating between Domains (DLID) [11]. This method learns multiple unsupervised deep models on the source, target, and mixed datasets, and uses the concatenation of features from every models as representation. In contrast, our work tackles the problem of cross-domain retrieval feature learning in the source and target domains simultaneously by designing a novel network architecture. Moreover, the traditional domain adaptation methods [11, 51] can even be applied on our learned features

for the refinement of retrieval results.



## Chapter 3

# Semantic Parts for Object Retrieval

In this chapter, we present an product retrieval system, involving object detection and feature embedding for image retrieval, to support a brand-new online shopping experience by taking footwear as concrete example. This system, called Circle & Search, enables users to naturally indicate any preferred objects by simply circling the products in images as visual queries, based on which the visually and semantically similar products are returned to users. The system is characterized by introducing *attributes* in both the detection and retrieval components. Specifically, we first develop an attribute-aware part-based detection model. By maintaining the consistency between parts and attributes, this detector has the ability to model high-order relations between parts, and thus enhances the detection performance. Based on the detection result, the system ranks all the objects in the repository using an attribute refinement retrieval model, which takes advantage of query-specific information and attributes correlation to provide an accurate and robust object retrieval. To evaluate this retrieval system, we build a large dataset with 17,151 footwear images, and each image is described by 10 semantic attributes, e.g., heel height, heel shape, sole shape, etc. In the experiment, the results indicate that the Circle & Search system achieves promising retrieval performance on footwear retrieval. As a general framework, we claim that this retrieval system can be easily extended into many other object domains for retrieval applications.

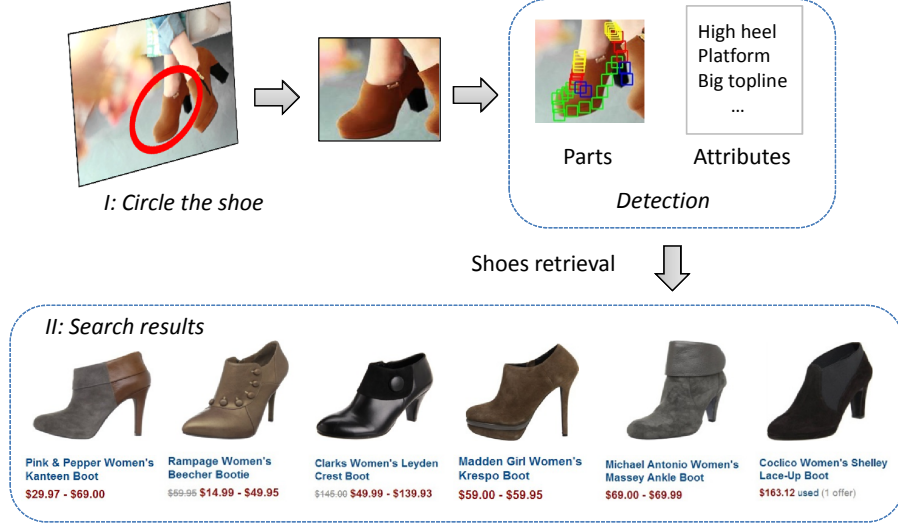


Figure 3.1: Scenario illustration of the Circle & Search footwear retrieval system. The user browses a website, and circles a shoe. The visually and semantically similar footwear images will be returned by the proposed system from the retrieval repository.

### 3.1 Introduction

Nowadays, the online shopping experiences are becoming increasingly important. Many websites, e.g., Amazon, eBay, Taobao, etc., provide convenient and economical platforms for people to buy their favorites. On those websites, the fashion-related commodities make a huge market, within which footwear takes a considerable proportion. However, some problems still exist when employing the current retrieval techniques on the online shopping websites. One of the most severe problems is the lack of semantic information in the product representation.

In this chapter, we propose an semantic footwear retrieval system, named Circle & Search. The application scenario of our retrieval system is illustrated in Figure 3.1. Imagining the scenario that the users browsing the online shopping websites are attracted by some products, which stimulates their purchasing desire. A common way to find the similar product is to type some proper keywords into the search engine and select their preference among the retrieval results. However, it is likely that one encounters difficulties when working out accurate descriptions as the keywords. One possible solution is to search the products with visual queries, that is, product images.

In our system, the visual queries can be the product images or user photos

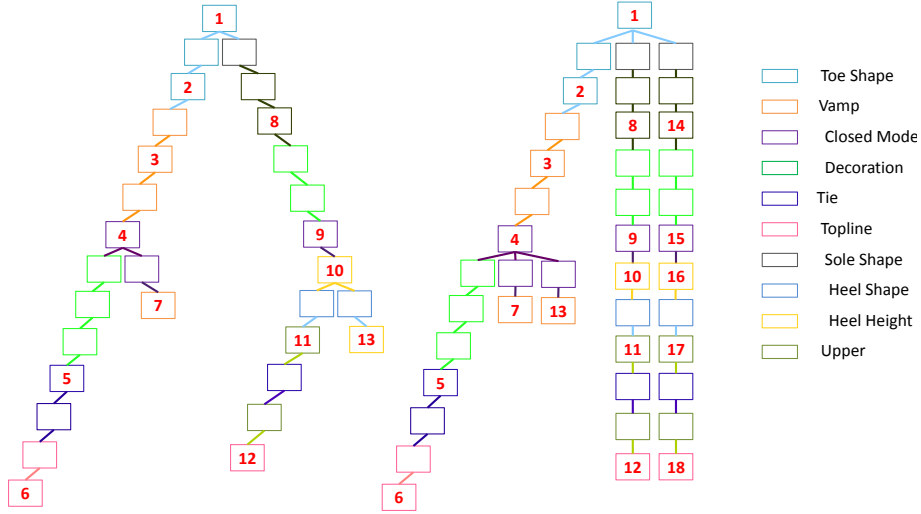
taken in daily life, which have large domain discrepancy. Particularly, the background of product image is relatively clean (like the retrieval result in Figure 3.1), while the background of daily photo is usually cluttered (like the query image in Figure 3.1). To decrease the domain discrepancy, previous studies demonstrate that the object parts are more expressive for the object representation than the whole image [5, 80]. By extracting the features from the semantic parts of objects, the influence of the background can be filtered out, and the discrepancy caused by viewpoints can be reduced [123]. Therefore, we carefully design some semantic parts of footwear for each viewpoint (see Figure 3.2 (a)), e.g., toe, heel, and vamp, etc, based on which some auxiliary parts are interpolated between every two manually defined parts for better representation (see Figure 3.2 (b)). To elaborately learn the spatial relations of parts, three tree-structure models are designed for three viewpoints, *i.e.*, the frontal, half profile and profile view, respectively.

Besides the object parts, many recent methods [8, 102, 115] propose that the semantic attributes can help to enhance the representation of objects. Traditionally, the attributes, e.g., high heel, round toe, etc., are used as the side information of visual features by attaching to the entire object. However, we propose that the attributes should be attached to certain *parts* of objects. For example, it is more reasonable that the attribute *toe shape* is combined with the *toe* parts. In this way, the attributes are used as the high-order relations among footwear parts in our system.

Generally, our system can be divided into the detection component and retrieval component. In the detection component, an attribute-aware part-based detection model is proposed. Specifically, several tree-structure models are designed to model the footwear in different views. The nodes of each tree represent the predefined footwear parts, and the edges between every two nodes indicate the deformation information between two parts. Contrast to the traditional framework, we claim that the key characteristic of our model is the using of consistency between attributes and relevant parts. Specifically, the appearance and deformation of parts should be influenced by the relevant attribute values, and vice versa. For example, if the attribute “toe shape” is of the value “round”,



(a) Manually labelled parts



(b) Tree-structures and correlations between parts & attributes

Figure 3.2: The semantic parts of footwear in three viewpoints are presented in (a). The color of each part indicates its corresponding attribute. The mapping of part color and attributes can be seen in the right part of (b). The semantic meaning of those parts is introduced in Sec 3.3.3. The number within each part indicates its order in the tree-structure model. By applying different combination strategies of manually labelled part in the detection model, we present two well-designed tree-structures for profile-view and frontal-view, respectively. Similarly, the color of each part indicates its corresponding attributes.

the visual appearance of the “toe” *parts* should be reasonably “round”. Meanwhile, the more precise location of *parts* can also help to improve the prediction of relevant attributes. Due to the consistency between parts and attributes, the detection of parts and the estimation of semantic attributes can be conducted via the Expectation-Maximization (EM) approach until the consistency is achieved.

In the retrieval component, the detected footwear parts and the predicted attributes are fed into a query-specific attribute refinement retrieval model. This refinement model aims to polish the attributes of query and retrieval images in the retrieval repository by taking advantage of correlation of attribute values. Based on the refined attributes, the footwear images in the repository are ranked and returned as retrieval result.

The main **contributions** of this work can be summarized as follows:

1. By defining the correlations between parts and attributes, we propose a novel attribute-aware part-based detection model. Due to the consistency between parts and attributes, the simultaneous footwear detection and attribute prediction can be performed efficiently in an EM manner.
2. The predicted attributes and semantic parts can explicitly explain the ranking criterion of our retrieval system. Therefore, the result of our retrieval system is more explicit and expressive. Moreover, our system can handle the non-rigid objects due to its part-based property, which is still the shortage of many state-of-the-art retrieval systems.
3. To our best knowledge, the query-specific attribute refinement retrieval model in this system is totally new in the retrieval area, which leads to a competitive retrieval result. To enable the practical applications of footwear retrieval, we collect a large-scale footwear dataset, with 17,151 images and 10 fine-grained attribute categories. Each footwear instance has  $3 \sim 4$  images of different viewpoints.

This capture is organized as follows. In Section 3.2, we briefly present the latest relevant research progress. In Section 3.3, the collection of footwear dataset is discussed. In Section 3.4, we introduce the Circle & Search system, including

the footwear detection model and footwear retrieval model. The experiments are demonstrated in Section 3.5. The concluding remarks of this capture are given in Section 3.6.

## 3.2 Related Work

In this section, we present the literature reviews on the object detection, image retrieval, and the exploitation of attributes.

**Object Retrieval :** The study of object retrieval has attracted many attentions both in academic and industrial field. Recently, a feedback system WhittleSearch [63] using the concept of “relative attributes” is proposed for the product retrieval in a user interaction manner. Besides the attribute, Lu et al. [84] harness the browse and search behaviours of users to improve the online shopping experience. This system is deployed on Tablet Pad by taking advantage of the multi-touch interfaces and user interactive behaviours. Shen et al. [100] proposed a method to automatically extract the object query for mobile product image search. By cropping the foreground object, the influence of cluttered background on visual features is removed, and the retrieval performance is significantly improved. Particularly, He et al. [47] proposed a novel mobile search system based on the “Bag of Hash Bits”, where the image is represented by a bag of hash bits. Generally, this system shows good searching performance and efficiency. Interestingly, Arandjelovic et al. [1] described a scalable method for smooth object retrieval, within which the real-time system can localize all the occurrences of outlined objects. By using the subspace decomposition, Jegou et al. [57] introduced a product quantization based method for approximate nearest neighbour search.

Referring to the practical applications, many websites provide the object search engine, e.g., Google Goggles<sup>1</sup>, Baidu Stu<sup>2</sup>, etc. which allow users to upload an image and return the similar products. Although no details are available about their techniques, it is likely that some visual features are extracted from the entire images or fixed patches of images, and certain distance metrics are

---

<sup>1</sup><http://www.google.com/mobile/goggles>

<sup>2</sup><http://stu.baidu.com>

designed to calculate the similarities.

**Object Detection:** Traditionally, the pictorial structure models [32, 23, 24] were proposed by using the mixtures of parts for object detection. However, those parts are greedily placed to cover the high-energy regions in a specific area. The unclear semantic meaning of parts results in the difficulty in associating the attributes with specific parts. In the problem of human pose estimation, Yang et al. [123] proposed an effective and flexible extension of part-based model. By defining different types of mixtures in each part, this model is feasible in many specially applications. For example, if type of mixtures is defined as the orientation of instance, the parts can precisely model the articulation of objects. Moreover, this model provides a general framework for modeling the co-occurrence relations of parts, as well as the spatial relations between parts, which construct the foundation of our detection model.

**Attributes Analysis:** The methods of attribute learning have been widely applied in many computer vision and multimedia tasks. Ferrari and Zisserman [35] presented a probabilistic generative model to learn the visual attributes. With the discriminative attributes [28], the objects are effectively categorized by using the compact attribute representation. Similarly, Parikh et al. [91] built a set of discriminative attributes by interactively displaying categorized object images to annotators and learn the attributes from annotators' feedback. Kumar et al. [68] proposed the attributes and simile classifiers to describe the face appearances, and conducted the competitive results for the application of face verification. Siddiquie et al. [102] explored the co-occurrence of attributes for image ranking and retrieval with multi-attribute queries. In the fashion-related area, researchers extracted the attributes by mining the images and their descriptive texts from the Internet [4], or by manually defining some domain-specific attributes [8] [80].

### 3.3 The Footwear Dataset

Recently, Kang et al. [61] collected a database of 5 million product images which contains 1.2 million objects in multiple viewpoints. Shen et al. [100] collected a



Figure 3.3: Some footwear examples of different viewpoints in our dataset. The images in the first row are the product images; The images in the second row are daily photos. Totally, our dataset contains 17,151 footwear images of several viewpoints.

dataset of sport product images in real-life with 43,953 images in 10 categories, including hats, shirts, trousers, shoes, socks, gloves, balls, bags, neckerchief and bands. However, the semantic information of object images are not collected in those datasets. In this chapter, we construct a new dataset specific for the footwear retrieval task with fine-grained attributes for our attribute-aware object retrieval problem.

### 3.3.1 Automatic Images Collection

Some example images of our dataset are shown in Figure 3.3. The images are collected from some online shopping websites (e.g., Amazon.com) and photo sharing websites (e.g., Flickr.com), by using the queries such as “shoes”, “footwear”, “boots”, and “sandals”, etc. Overall, 17,151 images are collected in this dataset.

### 3.3.2 Fine-grained Attributes

In this dataset, 10 footwear attribute categories are obtained. Those attributes are learned from the summaries of several online shopping websites. Some annotators are hired to select the correct attributes for each footwear instance in the dataset. To ensure the annotation accuracy, three annotators were assigned for each instance. A label is considered as correct if more than two similar annotations are selected. The illustration of footwear attributes, including the category and values, are shown in Figure 3.4.








































Attribute Name	Attribute Values			
Heel Shape				
	High-thin	Thick	Cubic	Wineglass
Sole Shape				
	Cone	Wedge	Flat	
Decoration				
	Frontal	Lateral	Rear	None
Tie				
	Platform	Thick	Flat	
Topline				
	Front-strap	Back-strap	None	
Toe Shape				
	Round	Square	Pointed	Fish-mouth
Vamp				
	Net	Dot	Stripe	Colorful
Upper				
	Super-high	High	Middle	Low
Closed Mode				
	Velcro	zipper	Shoelace	Elastic
Heel Height				
	Super-high	High	Middle	Short

Figure 3.4: The illustration of footwear attributes. The first column of each table is the attributes category, and the rest columns of each table are the corresponding attribute values.

### 3.3.3 Semantic Footwear Parts

To handle the out-of-plane rotation, three views are defined according to the rotation angle of footwear instance. Generally, the angles of images in frontal, half profile, and profile view are around  $0^\circ$ ,  $45^\circ$ , and  $90^\circ$ , respectively. For each image, 13 parts are defined for the footwear in profile and half-profile view, within which 11 parts are selected for the footwear in frontal view. The same annotators are hired to label the views and parts of footwear instances, and each image is annotated by three annotators. The images whose views are not agreed by all three annotators are filtered out. Finally, the average positions of parts are used as the ground-truth part annotation if the viewpoints of footwear instances are labelled.

To fully capture the visual appearance of footwear in each viewpoint, we design several tree-structure models, and select the optimal structure for each viewpoint by applying the deformable model [123] on each structure. The structures of model for the profile and frontal viewpoints are shown in Figure 3.2 (b). The nodes of each tree model represent the footwear parts, and the edges are the deformation information between every two parts. The semantic attributes

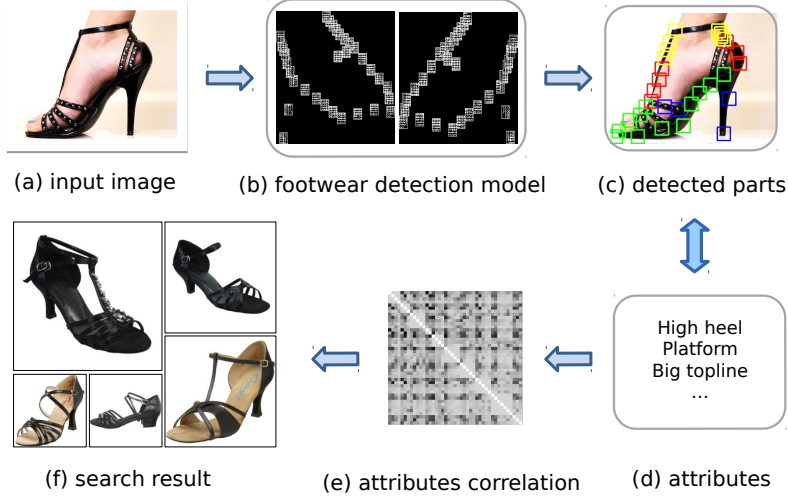


Figure 3.5: The framework of our proposed footwear retrieval system: given a footwear image (a), the parts (c) and attributes (d) of the footwear instance are predicted by our attribute-aware part-based detection model (b) in an iterative manner. These attributes are then fed into a query-specific attribute refinement retrieval model (e) for refinement. (f) The refined attributes are used as retrieval feature to retrieve the retrieval results.

are attached to several relevant parts. Note that footwear instances of different viewpoints contain different number of parts.

### 3.4 The Circle & Search System

The framework of our system is illustrated in Figure 3.5. Given a query image, the attribute-aware part-based detection model detects the locations of footwear parts and predicts the corresponding attributes. Those predicted attributes, which fully capture the properties of query image, are concatenated to form the semantic feature. Based on this feature, the retrieval images are ranked and returned as retrieval result. However, because different attributes are predicted independently, there should be noise in the predicted result. To refine the retrieval result, we utilize the co-occurrence of attributes in the retrieval repository to refine the predicted attributes by our query-specific attribute refinement model. Finally, the retrieval images are ranked and returned by using the refined attributes as retrieval feature.

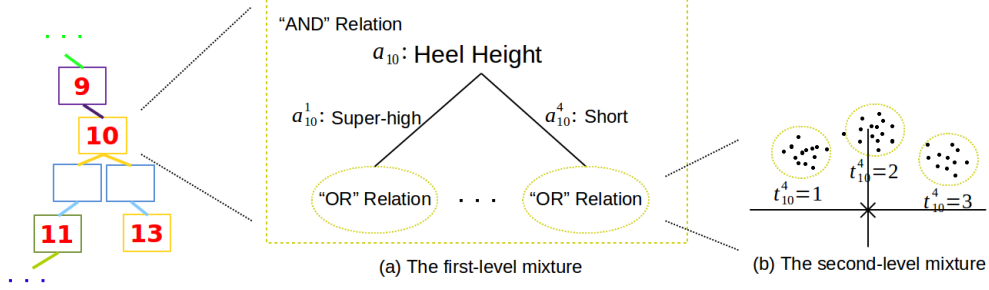


Figure 3.6: The sub-figure (a) presents the “AND-OR” hierarchical structure of 10-th part. Firstly, the 10-th part of training samples are divided into several components with “AND” relation defined by attribute values, which constructs the first-level mixture (a). Each component of first-level mixture is consisted of  $K$  components with “OR” relation generated by  $K$ -means (b). The cross at the origin of coordinates in sub-figure (b) represents the position of 10-th part’s parent node, *i.e.*, 9-th part, and the features of parts for  $K$ -means are the normalized distance along the x-axis and y-axis from 10-th part to 9-th part.

### 3.4.1 Attribute-aware Detection

In this section, we elaborate the details our proposed attribute-aware detection model. Specifically, the part of this model is carefully designed by embedding attributes into each part to construct the hierarchical mixture. Based on this mixture part, the detection model is discussed in three aspects, *i.e.*, appearance model, deformation model, and attribute integration. Finally, the inference and learning of this model are introduced in this section.

#### Hierarchical Mixture

Inspired by the part-based detection approaches [32, 123], the tree-structure models constructed by a set of deformable semantic parts are used to represent the footwear instances of different views. Due to the variant types of footwear, the visual appearance of footwear instance can not be fully captured by the low-level image features. For better representation, the footwear attributes are introduced into our detection model.

Denote the footwear image as  $I$ , and  $\{l_i = (x_i, y_i)\}_{i=1}^N$  are the positions of  $N$  footwear parts in  $I$ , and  $\{\alpha_i\}_{i=1}^M$  is the  $M$  footwear attributes. Usually, each attribute has several attribute values (see Figure 3.4), and we write  $\alpha_i = [\alpha_i^1, \dots, \alpha_i^{n_{\alpha_i}}]$ , where  $n_{\alpha_i}$  is the number of values of attribute  $\alpha_i$ , and each element

in  $a_i$  is the probability of a footwear instance having the corresponding attribute value. In our model, one attribute are associated with several parts, and we simplify the model by assuming that each part only attaches to one attribute. Intuitively, the attributes maintain the high-order relations between parts. The mapping between footwear parts and attributes can be obtained by defining two functions, *i.e.*,  $i_a = f_a(i_p)$  and  $\mathbf{i}_p = f_p(i_a)$ , where  $\mathbf{i}_p$  denotes the indices of footwear parts affected by the  $i_a$ -th attribute, and  $i_p \in \mathbf{i}_p$  is the index of one specific part. Given the index of specific part  $i_p$ , we can easily find out its corresponding attribute by  $f_a(i_p)$ .

For notational convenience, the attribute of the  $i$ -th part will be simply denoted as  $a_i$  as long as no confusion is caused, and similarly we define  $a_i = [a_i^1, \dots, a_i^{n_{a_i}}]$ . With such notation, we should notice that  $a_i$  and  $a_j$  may be the same attribute, as one attribute is associated with several parts.

Intuitively, the attribute  $a_i$  attached to the  $i$ -th part has an explicit effect on the visual appearance of this part. In other words, the same part in different images may be combined with different attribute values (e.g., the attribute *toe shape* may be *round*, or *pointed* in different images), which thus results in the discrepancy of appearance of this part in different images. Therefore, a first-level mixture, which is discriminated by the values of attribute  $a_i$ , is constructed for  $i$ -th part (see Figure 3.6 (a)). Obviously, the number of attribute values  $n_{a_i}$  is the number of components in the first-level mixture of  $i$ -th part. Note that the relation of components in first-level mixture is modelled as “AND” in our model, as we suggest that each value of specific attribute is softly assigned to its related parts with certain probability.

Though the first-level mixture can considerably reduce the appearance discrepancy of the same parts between different images, we argue that the displacement between two parts is still different among footwear instances, even if the same part in different instance images have the same attribute value. To model the deformation of parts, the same parts in different images with the same attribute value are clustered into several groups according to their normalized spatial distribution.

Specifically, the deformation type of  $i$ -th part with attribute value  $a_i^k$ , denoted

by  $t_i^k \in \{1, \dots, K\}$ , is integrated into each component of first-level mixture to construct a second-level mixture (see Figure 3.6 (b)). Note that  $K$  is the number of components in each second-level mixture, and we make  $K$  the same in our detection model. To generate the second-level mixture,  $K$ -means is applied on the training samples by using the normalized distance between each part and its parent part along the x-axis and y-axis as feature. Obviously, the centroids of  $K$ -means stand for the components of second-level mixture. Different from the relations of components in first-level mixture, the relation of deformation types  $t_i^k$ , *i.e.*, the components of second-level mixture, is modelled as “OR” node. In other words,  $t_i^k$  can only be exclusively selected as one integer within the value range  $\{1, \dots, K\}$ .

Therefore, the  $t_i = [t_i^1, \dots, t_i^{n_{a_i}}]$ , which is the types of components in the first-level mixture in  $i$ -th part, can represent the hierarchical “AND-OR” structure of  $i$ -th part. This hierarchical mixtures can significantly stabilize the performance of our tree model, as it can greatly reduce the discrepancy resulted by the visual appearance and the deformation of parts.

### Model Formulation

To introduce the model, let's denote  $G = (V, E)$  as single tree, where  $V$  is the set of tree nodes, with each node correlated to a footwear part, and  $E$  is the set of tree edges. The score function  $S(I, a, t, l)$  of this tree model can be written as follows with the configuration of attributes  $a$ , part types  $t$ , and positions  $l$ :

$$S(I, a, t, l) = \sum_{i \in V} a_i \odot (w_i^{t_i} \odot \phi(I, l_i) + b_i^{t_i}) + \sum_{ij \in E} a_{ij} \odot (w_{ij}^{t_i, t_j} \odot \theta(l_i, l_j) + b_{ij}^{t_i, t_j}), \quad (3.1)$$

where  $\phi(I, l_i) \in \mathbb{R}^d$  is the HoG feature of image  $I$  extracted at  $i$ -th part, and  $\theta(l_i, l_j) = [dx, dy, dx^2, dy^2]^T \in \mathbb{R}^4$  indicates the relative position of the  $i$ -th part to  $j$ -th part by defining  $dx = x_i - x_j$  and  $dy = y_i - y_j$ .  $a_i \in \mathbb{R}^{n_{a_i}}$  is the probability vector of  $i$ -th part containing the values of attribute  $a_i$ . This probability vector is used to model the influence of attribute  $a_i$  on this part. Similarly,  $a_{ij} \in \mathbb{R}^{n_{a_i}} \times \mathbb{R}^{n_{a_j}}$  is the joint attribute value probability matrix of the  $i$ -th part and  $j$ -th part,

which is used to model the effect of pairwise attributes on two adjacent parts.  $\omega_i^{t_i} \in \mathbb{R}^{n_{a_i} \times d}$  and  $\omega_i^{t_i, t_j} \in \mathbb{R}^{n_{a_i} \times n_{a_j} \times 4}$  are the model parameters to be learned. Note that here  $\odot$  is a generalized dot product operator which can be performed on two tensors of different orders and dimensions. Denoting  $A \in \mathbb{R}^{m_1 \times \dots \times m_p \times n_1 \times \dots \times n_q}$  and  $B \in \mathbb{R}^{n_1 \times \dots \times n_q}$ , each element in the result tensor  $C = A \odot B \in \mathbb{R}^{m_1 \times \dots \times m_p}$  is calculated as:

$$C(i_1, \dots, i_p) = \sum_{j_1} \dots \sum_{j_q} A(\dots, j_1, \dots, j_q) B(j_1, \dots, j_q). \quad (3.2)$$

**Appearance Model:** The first term in Equation (3.1), called as appearance model, indicates the local response of putting a set of templates  $w_i^{t_i}$  at position  $l_i$  for the  $i$ -th part by tuning the attribute value probability vector  $a_i$  and the types  $t_i$ . It should be emphasized that the types  $t_i \in R^{n_{a_i}}$ , and each element  $t_i^{k_i}$  in the type vector  $t_i$  indicates the index of component in the  $k_i$ -th second-level mixture of  $i$ -th part. The bias  $b_i^{t_i}$  is the preference of assigning types  $t_i$  to the  $i$ -th part with different attribute values. Obviously, the formula of appearance model indicates its “AND-OR” node structure, as only one response of template from each second-level mixture is selected, and the response of appearance model in each part is the weighted sum of selected response from every second-level mixture.

**Deformation Model:** Given a specific combination of joint attribute values  $(a_i^{k_i}, a_j^{k_j})$  for the adjacent  $i$ -th part and  $j$ -th part, the different combinations of displacement  $(t_i^{k_i}, t_j^{k_j})$  are determined by the normalized distance between these two parts. Each combination presents the particular relative placement of  $i$ -th and  $j$ -th part under the condition that  $i$ -th (or  $j$ -th) part is assigned as  $k_i$ -th (or  $k_j$ -th) attribute value with probability  $a_i^{k_i}$  (or  $a_j^{k_j}$ ). By tuning the combinations of types for two adjacent parts and the probability vector of their joint attribute values, the second term in Equation (3.1), also known as deformation cost, controls the co-occurrence of spatial information and joint attributes combination between two parts. Similarly, the bias  $b_{ij}^{t_i, t_j}$  presents the preference of particular co-occurrence of types combination  $(t_i, t_j)$ .

**Attribute Integration:** In our detection model, we suggest that the se-

semantic attributes can affect the visual appearance of multiple parts and the deformation between every two parts. Therefore, the attributes are considered as high-order relation of relevant parts. For example, if the attribute *heel shape* of certain shoe is of the value *high-thin*, the two ending parts of heel should be relatively *thin* at the same time. Therefore, the selection of optimal templates for parts will be constrained to fit the global relations preserved by attributes. To integrate the attribute into one part, we rescale the response of templates in each second-level mixture by multiplying the probabilities of corresponding attribute values, and sum up the rescaled response of optimal template from each second-level mixture as the response of this part.

Specifically, attribute classifiers are pre-trained by using the concatenated low-level image features of detected parts. During the detection, the probability vector of attribute  $a_i$  can be obtained by inputting the concatenated features of current detected parts into corresponding classifier. Each element of probability vector  $a_i^k$  indicates the probability of relevant parts having the  $k$ -th value of attribute  $a_i$ , *i.e.* the  $k$ -th component of first-level mixture in  $i$ -th part. For each component in the first-level mixture, its response is the highest score of template in the second-level mixture multiplies its attribute value probability. This again indicates that the displacement type in each second-level mixture is selected as “OR” relation, and the component in every first-level mixture is selected as “AND” relation. Note that one specific attribute may affect several parts at the same time, which indicates that the placement of parts must be effected by the high-order relation maintained by attributes. Therefore, the hierarchical “AND-OR” structure and high-order relations can enhance the performance of part detection and attribute prediction by unifying the appearance of parts and the semantic meaning of attributes.

## Model Inference

With the learned model parameters  $(\omega_i, \omega_{ij}, b_i, b_{ij})$ , we can detect the parts and attributes of footwear by maximizing the score function  $S(I, a, t, l)$  over  $a$ ,  $t$ , and  $l$ . In practice, the feature pyramid is extracted to decide the optimal scale. However, with the integration of attributes, the score function  $S(I, a, t, l)$  is hard

to solve due to its non-convex property. Fortunately, EM-based approach can be applied in the inference to iteratively achieve the solution. Generally, when fixing  $a$ , Equation (3.1) becomes convex over  $t$  and  $l$ , and can be effectively solved by dynamic programming due to its tree-structure [32, 123]. After getting the current optimal  $t$  and  $l$ , we can calculate the expected attributes  $a$  according to the current predicted parts. These two steps iterate until convergence is achieved.

**Initialization:** For notational convenience, we define  $z_i = (l_i, t_i)$  as the location and types of  $i$ -th part. In the initialization, we aim to detect the initial parts of footwear instance. To eliminate the influence of attributes, we pre-trained a normal detection model without attributes, and get an initial estimation of the locations and types of parts via this detection model.

**E-Step:** The E-Step aims to estimate the expected attributes  $a$  based on the positions  $l$  and types  $t$ . Specifically, the relations between parts and attributes are predefined by training a multi-class linear SVM for each attribute. The input of SVM is the concatenation of low-level image features extracted from relevant parts. Based on the detected parts in the previous step (denoted as  $z$ ), we concatenate the features of corresponding parts and estimate the expectation values of attributes from the normalized scores estimated by the multi-class linear SVM. Formally, this procedure can be written as:

$$\hat{a}_i = f_i(\varphi(I, z_{f_p(f_a(i))})), \quad (3.3)$$

where  $f_i(\cdot)$  denotes the attribute classifier. The  $\varphi(I, z)$  denotes the concatenated SIFT feature of  $i$ -th part at the position  $l_i$  of image  $I$ , where  $l_i \in z$ .

**M-Step:** The M-Step aims to estimate the position  $l$  and type  $t$  in  $z$  of every part based on the updated attributes  $a$ . This procedure can be conducted by fixing attributes  $a$  and evaluating the following objective function using dynamic programming,

$$\text{score}_i(z_i) = \hat{a}_i \odot (w_i^{t_i} \odot \phi(I, l_i) + b_i^{t_i}) + \sum_{k \in \text{kids}(i)} m_k(z_i), \quad (3.4)$$



$$m_k(z_i) = \max_{z_k} [\text{score}_k(z_k) + \hat{a}_{ki} \cdot (w_{ki}^{t_k, t_i} \cdot \theta(l_k - l_i) + b_{ki}^{t_k, t_i})], \quad (3.5)$$

where  $\text{kids}(i)$  is the set of children nodes of  $i$ -th part, and it is empty for the leaf parts. During detection, the tree model starts from the leaves and moves upwards until arriving at the root node.

Given a fixed attribute  $\hat{a}_i$ , Equation (3.4) calculates the current local score of the  $i$ -th part over every pixel position  $l_i$  and types  $t_i$ , and collects the score message from its children. Particularly, the first term in Equation (3.4) can be computed by traversing the whole feature pyramid with different types of mixtures. Equation (3.5) adds the local score of the  $k$ -th part with relative deformation cost, and passes the best score as message to its parent node. Note that the fixed pairwise attribute  $\hat{a}_{ki}$  is computed in the previous expectation step. Once the message arrives at the root part, the configuration of the root part with the best score becomes the optimal configuration of current detection over position  $l_1$  and type  $t_1$ . By keeping the trace of message passing, one can back track the direction from root to leaves to get the optimal configuration of each part.

**Time Complexity:** Due to the linear time complexity of E-Step, the time complexity of our model is mainly decided by the complexity of M-Step. The M-Step concentrates on the dynamic programming in Equation (3.5) over every positions  $l$  and types  $t$ . In practice, the distance transform [29] is used to calculate the message of each part on every candidate positions  $l$  with  $O(|l|)$  complexity. By looping over every  $|t| \times |t|$  possible types of parent nodes and children nodes, the complexity of this part becomes  $O(|l| \times |t| \times |t|)$ , which is also the complexity of our detection model.

### Model Parameter Learning

The model is trained in a supervised learning paradigm. Given the labelled positive example set  $I_{\text{pos}}$  with annotations  $(a_{\text{pos}}, t_{\text{pos}}, l_{\text{pos}})$  and negative example set  $I_{\text{neg}}$ , we aim to solve a structured object function similar to those proposed in [32, 123]. For notational convenience, we denote  $y_n = (a_n, t_n, l_n)$  as the prior information, where  $n \in \text{pos}$ . Since the score function is linear in model parame-

ters  $\beta = (\omega, b)$ , it can be re-written as  $S(I, y_n) = \beta \cdot \Phi(I_n, y_n)$ , where  $\Phi(I_n, y_n)$  is the concatenation of appearance and deformation features. To maximize the score function, the model is learned in the max-margin form:

$$\begin{aligned}
& \min_{\omega, \xi_n \geq 0} \frac{1}{2} \beta \cdot \beta + \lambda \sum_n \xi_n \\
& \text{s.t. } \forall n \in \text{pos}, \quad \beta \cdot \Phi(I_n, y_n) \geq 1 - \xi_n \\
& \quad \forall n \in \text{neg}, \forall y, \quad \beta \cdot \Phi(I_n, y) \leq -1 + \xi_n
\end{aligned} \tag{3.6}$$

The above quadratic program problem is known as structural SVM and can be solved by many optimization solvers. In our experiments, the dual coordinate descent algorithm developed in [123] is adopted to solve the problem. The learning procedure can be roughly separated into two sub-procedures. The first sub-procedure includes the construction of hierarchical mixtures, the learning of separate mixtures, and the learning of part deformations with the labelled attribute values. The second sub-procedure is the adjustment of parts and attributes, as the consistency of parts and attributes is also required in the learning procedure.

**Learning with labelled Attributes:** As aforementioned, each part in the tree model is consisted of a hierarchical mixture, where the first-level mixture discriminated by the labelled attribute values. In each component of first-level mixture, the second-level mixture is constructed by  $k$ -means with the relative distance between parent part and children part as feature. At the beginning, the template of each component in the second-level mixture in one part is trained on the image parts which contain the corresponding attribute value and type. This indicates that the element of attribute probability vector  $a_i$  of the  $i$ -th part is one if that element is the labelled as attribute value. After learning the templates for each part, the weights of deformation model are calculated in a second round training with labelled information. Consequently, an initial tree model is constructed. Note that the attribute classifiers are also learned separately according to the corresponding features and labelled attributes.

**Adjustment of Parts and Attributes:** The detection problem is easy if

the labelled attributes are available both in the training and testing procedures. However, the problem becomes tough when the attributes are unknown in the testing procedure. In the inference section, we introduce an EM-based approach to keep the consistency between parts and attributes by choosing the optimal mixtures for every part. However, such consistency may not be achieved unless the tree model and attribute classifiers are consistent in the training procedure as well. Therefore, the EM-based approach is also conducted in the learning procedure. Note that the consistency in the training procedure is achieved by adjusting the weights of templates for each part, rather than selecting the optimal template.

In this sub-procedure, the parts and attributes of the training images are slightly adjusted, and thus the attribute classifiers and tree model will be updated accordingly. Specifically, assuming that the tree model  $tree^{(i)}$  is trained by the image parts  $l^{(i)}$ , types  $t^{(i)}$  and attributes  $a^{(i)}$  at the  $i$ -th step, we can re-detect the parts  $l^{(i+1)}$  on every training image, and thus get the attributes  $a^{(i+1)}$  according to the hierarchical structure. Based on the detected parts  $l^{(i+1)}$ , we can predict their attributes  $a^{(i+1)'}$  by the pre-trained classifiers  $classifier^{(i)}$ . Then, the attribute classifiers are re-trained as  $classifier^{(i+1)}$  by the parts  $l^{(i+1)}$  and attributes  $a^{(i+1)}$  from tree model  $tree^{(i)}$ . Meanwhile, the tree model can also be updated as  $tree^{(i+1)}$  by the new attributes  $a^{(i+1)'}$  from attribute classifiers. Those two steps are iterated until convergence. In the experiments, we find the training is converged only in two or three iterations. Similar findings are also reported by [123].

### 3.4.2 Query-Specific Attribute Refinement

The probability vector of attribute values of query image, denoted by  $x$ , can be calculated by the detection model introduced in Section 3.4.1. The task of footwear retrieval is to calculate the similarity between  $x$  and the attribute value probability vector  $y$  of each candidate image in our retrieval repository  $Y$ . Traditionally, the ranking criterion uses the Euclidean distance, *i.e.*,  $\|x - y\|_2$ .

However, the attribute value probability vectors  $x$  and  $y$  may be noisy. We propose to refine them by considering the correlations between different attribute

values. For example, the value “*high heel*” of attribute “*heel height*” usually appears with the the value “*fish mouth*” of attribute “*toe shape*”. To model the pairwise correlations between different attribute values, we obtain the co-occurrence matrix  $C$  of the attributes from the training dataset. Then, we calculate the Laplacian matrix  $L$  based on the co-occurrence matrix  $C$ . In this work, we propose the query-specific attribute refinement method by optimizing the following function:

$$\min_{x', y'} \|x' - x\|_2^2 + \|y' - y\|_2^2 + \alpha(x'^T L x' + y'^T L y'), \quad (3.7)$$

where  $x'$  and  $y'$  are the refined attribute values probability vector of  $x$  and  $y$ , respectively. The first two terms require that the refined attribute values should be similar to the original attribute values. The third term requires that the refined attribute values should follow the correlations of attributes. The underlying intuition is that Equation (3.7) aims to align  $x$  and  $y$  by removing the possible noises existing in  $x$  and  $y$ . Obviously, Equation (3.7) can be re-formalized as:

$$\min_{x', y'} \begin{bmatrix} x' \\ y' \end{bmatrix}^T D \begin{bmatrix} x' \\ y' \end{bmatrix} - 2 \begin{bmatrix} x \\ y \end{bmatrix}^T \begin{bmatrix} x' \\ y' \end{bmatrix}, \quad (3.8)$$

where  $D = \begin{bmatrix} I + \alpha L & 0 \\ 0 & I + \alpha L \end{bmatrix}$ , and Equation (3.8) can be solved by setting the derivative as zeros, and thus:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = D^{-1} \begin{bmatrix} x \\ y \end{bmatrix}. \quad (3.9)$$

Based on the refined attribute from Equation (3.9), we can get the final ranking according to  $\|x' - y'\|_2$ .

### 3.4.3 Extension of Retrieval System

Generally, our Circle & Search system contains two sub-modules, *i.e.*, attribute-aware part-based detection model and query-specific attribute refinement re-

trieval model. Though we take footwear as concrete example to introduce our retrieval system, we suggest that our system can be easily extended into other domains as long as the semantic attributes of objects in those domains can be considered as high-order relations between parts.

Taking clothes as example, we can first define a human-shape model to represent the clothes, including upper body and lower body. Meanwhile, the attributes of clothes can be collected from some professional websites, e.g., collar, sleeve, and shape, etc. (please refer to Chapter 5 for more clothing attributes). With those information, one attribute may affect the appearance of several relevant parts, which is known as high-order relation in our detection model. For example, the V-shape collar requires that the part near the chest should be “V” shape, and the parts near both shoulders should be straight; While, the round collar requires that the parts beside the chest should be relatively round. Therefore, we can easily train an attribute-aware part-based clothes detection model if the attributes of clothes indeed affect the appearance of several parts. With the result from our detection model, the query-specific attribute refinement model can calculate the co-occurrence matrix of clothes attributes, and apply the attribute refinement method on the predicted attributes of query clothes and the ground-truth attributes of clothes in the repository. Finally, the similar clothes in the repository should be found according to the ranking of refined attributes similarity.

### 3.5 Experiments

In this section, we evaluate the performance of our Circle & Search system on our collected database in terms of footwear detection and retrieval. Our footwear database contains 17,151 footwear images, including product images and daily life photos. In the detection experiment, the detection of parts and prediction of attributes are used to evaluate the performance of detection models. In the retrieval experiment, the query-specific attribute refinement model refines the predicted attributes, and ranks the retrieval images according to the refined attributes.

Generally, the detection result indicates that our attribute-aware part-based detection model improves the performance of part detection and attribute prediction. In the retrieval experiment, we claim that the query-specific attribute refinement based approach leads to a promising performance on cross-domain footwear retrieval problem.

### 3.5.1 Exp1: Semantic Footwear Detection

**Experimental Settings:** In this experiment, we compare our detection model with the flexible mixtures-of-parts model proposed in [123]. Similar to our model, we apply the baseline model on several pre-defined tree-structures, and select the tree-structure with the best performance for each viewpoint. For fair comparison, the configuration of baseline and our model are almost the same, except that the number of components in each mixture, *i.e.*,  $K$ , is 6 in baseline, while  $K$  is 3 in our model. This indicates that we give prior advantages to the baseline. Beside the manually defined parts, auxiliary parts are interpolated between two labelled parts to enrich the representation of footwear instance in specific viewpoint for both baseline and our model.

In the training procedure, 750 images (250 images for each view), including product images and daily life photos, are carefully selected from our footwear database as positive training images. Note that the rotation and flip operation are performed on the training data for data augmentation, which creates 10 times training samples. The INRIA database [14] is used as our negative training set. The negative mining strategy is applied on the selection of negative samples. To evaluate the detection performance, 2,250 images (750 images for each view) consisting of product images and daily life photos are used as the test dataset.

We conduct two experiments to comprehensively evaluate our detection model. In the first part of detection experiment, we assume that the viewpoints of testing images are given, and the detector for corresponding viewpoint is applied for detection. In the second part of detection experiment, the viewpoints of testing images are unknown. To obtain the viewpoints, the normalized predicted scores are introduced, so that the scores among three models are comparable.

To evaluate the performance of detection, the metrics Average Precision of

Table 3.1: The comparison of detection performance between baseline and our method in different settings.

Detection Model	Viewpoint	Mean APK	Mean PCK
Part-based Model [123]	Frontal View	48.7%	64.8%
	Half Profile View	60.3%	72.6%
	Profile View	59.5%	73.4%
	Unknown	53.5%	66.7%
Our Detection Model	Frontal View	50.3%	66.5%
	Half Profile View	63.4%	75.8%
	Profile View	64.3%	76.8%
	Unknown	57.9%	70.5%

Keypoints (APK) and Precision of Correct Keypoints (PCK) [123], are employed to evaluate the detection of parts. For the APK, the candidate is considered to be correct (true positive) if it lies beside the ground truth part. Particularly, this metric can correctly penalize both missing-detection and false positives. The PCK evaluation explicitly factors our detection by requiring the testing images to be annotated with tightly-cropped bounding box for each footwear instance. Note that we directly consider the images with wrong predicted viewpoints as the incorrect prediction in the second part of detection experiment.

In the attribute prediction, because the baseline cannot predict the attributes, we use the multi-class linear SVM to predict the attributes by extracting the SIFT features from the circled images and parts detected by [123], and the ground-truth parts, respectively. The precision is used to evaluate the performance of attribute prediction.

**Detection Performance:** The results of part detection are demonstrated in Table 3.1. Generally, our attribute-aware part-based detection model outperforms the baseline in both settings. On average, the mean APK and mean PCK of our model are about 3.17% and 2.77% higher than that of baseline if the prior knowledge of viewpoint is given. This indicates that the integration of attributes is helpful on the improvement of detection performance. However, compared with the results in profile and half profile viewpoints, the improvement of model performance in the frontal viewpoint is slightly lower. A possible explanation is that some distinctive parts of footwear in frontal viewpoint are self-occluded.

In the second part of detection experiment, because the viewpoints of testing

images are unknown, the detection models need to predict the viewpoints of the testing images. To make the scores of three models comparable, we normalize the scores of three models according to the score distributions on training images predicted by their corresponding model. After normalization, the viewpoint with highest score is considered as the predicted viewpoint of each image. Using this normalization strategy, the prediction accuracy of viewpoint can reach up to 97.1% and 95.4% in our model and the baseline [123], respectively. After obtaining the predicted viewpoint, the detected parts and attributes in the predicted viewpoint are regarded as the detection result. Generally, compared with the first part of detection experiment, the performance of both detection models in the second part decreases slightly. Obviously, this is mainly due to the scarcity of viewpoint information. However, the mean APK and mean PCK of our detection model are still about 4.4% and 3.8% higher than baseline. Compared with the improvement in the first detection experiment, this indicates that our model is more stable than the baseline if the viewpoints of testing images are unknown.

For the running time, our model costs about twice amount of time than the baseline. Specifically, our model spends about 2.5 seconds on processing a typical footwear image with  $500 \times 500$  pixels resolution, and the part-based model [123] costs about 1.5 second for each image.

In the experiment of attributes prediction, by extracting SIFT feature from the circled images, detected parts of Part-based Model [123], and the ground-truth parts, we implement three baselines with multi-class linear SVM. Particularly, as the huge efforts are needed to circle every testing images, we directly use the ground truth parts to generate the foreground bounding boxes as the circled images. Obviously, the third baseline with ground-truth parts reasonably indicates the upper-bound performance of attribute prediction. To compare with our result, the inputs of three baselines are twofold: the parts from image of specific viewpoint, and the parts from image with unknown viewpoint.

Table 3.2 presents the precision of attributes predicted by three baselines and our model. In average, if the viewpoints of the testing images are available, the precision of our model is about 14.12% and 7.42% higher than the first two



Table 3.2: Attribute classification accuracy of three baselines and our method.

View	Bounding box	Parts from [123]	Our model	Upper bound
Frontal View	64.30%	71.77%	79.90%	82.69%
Half Profile View	67.65%	73.77%	80.73%	85.46%
Profile View	68.12%	74.62%	81.80%	85.89%
Unknown	62.22%	70.89%	75.78%	80.04%

baselines, and it is about 3.87% lower than the upper-bound. If the viewpoints of parts are unknown, our prediction precision is about 13.56% and 4.89% higher than the baselines, and it is about 4.26% lower than the upper-bound. Overall, the low precision of the baseline using bounding box may be caused by the cluttered background of testing images. Compared with the baseline using parts of Part-based Model [123], we conclude that the attributes and visual representation of parts can enhance each other in our model. Moreover, the gap between the baseline using ground-truth parts and our model is relatively smaller than the gaps between the other baselines and our model. We suppose that the re-adjustment strategy which uses the consistency of parts and attributes in our detection model contributes to the reduction of performance gap.

**Examples of Footwear Detection:** To demonstrate the result of our footwear detection model, we present some testing examples of product images and daily photos returned by our detection model in Figure 3.7. The results illustrate that our model achieves good performance, especially in the vamp part, sole part, and heel part. However, the detection performance of upper parts still needs to be improved. The imprecision of these parts is due to the significant discrepancy of footwear upper between high-upper shoe and low-upper shoe, such as *boots* and *sports shoes*.

Observing from the experiment, we claim that using the parts to represent footwear can greatly reduce the noise caused by cluttered background. Meanwhile, by using the constraint between attributes and parts, we can get appreciable improvement both in part detection and attribute prediction.

### 3.5.2 Exp2: Attribute-aware Footwear Retrieval

In this subsection, we comprehensively evaluate the performance of our retrieval system, *i.e.*, the query-specific attribute refinement retrieval system, by com-

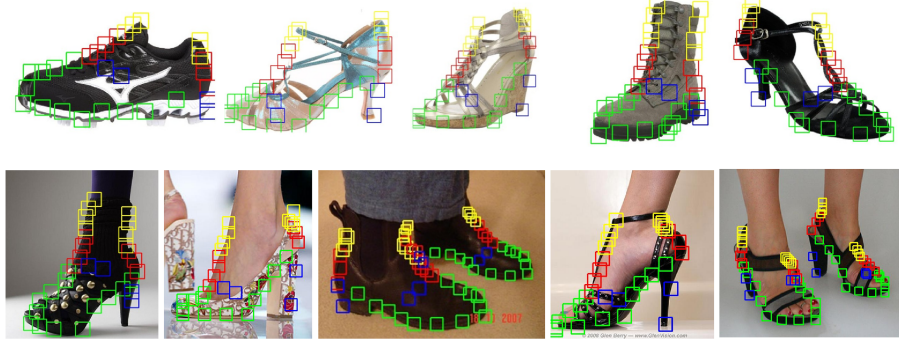


Figure 3.7: Some examples of the detected bounding boxes. In the result, we can observe that our detection model can effectively localize the different footwear parts even the scale, viewpoint are quite diverse or background is cluttered.

paring with variant baselines, including the state-of-the-art retrieval systems. The experiment result presents that our retrieval system outperforms the other baselines by using the refined attributes as semantic features.

**Experimental Settings** In the retrieval experiment, 200 product images and daily photos are used as query images and the rest product images construct the retrieval repository. To make the experiment convincing, three searching strategies are used. The first searching strategy is implemented by using the similarities of SIFT feature between query images and images in retrieval repository. The second strategy is implemented by using the similarities of predicted attribute probabilities between query images and images in retrieval repository. The last searching strategy is our proposed retrieval method, *i.e.*, the query-specific attribute refinement retrieval method, which uses the refined attributes for retrieval.

To evaluate the effectiveness of our detection model, we use the parts and attributes of three methods mentioned in detection experiment as the input for three retrieval strategies. Specifically, those three data sources are the result of multi-class linear SVM on cropped footwear image, the result of Part-based Model [123], and the result of our detection model. By using the parts and attributes from one of the three aforementioned methods, we can compare the performance of the three retrieval strategies. By using specific retrieval method, we can evaluate the effectiveness of our detection model on the retrieval performance. Note that to guarantee the fairness of comparison, the 5-folder cross-

validation is used in our experiment.

To further evaluate our retrieval model, two state-of-the-art object retrieval methods, *i.e.*, BoB with segmentation method [1] and BoHB with PCA Hashing strategy (r=2) and boundary re-ranking [47], are compared in this experiment. The configuration of training data for these two baselines is similar to our detection experiment. Specifically, 750 images are used as training images for super-pixel classifier in BoB with segmentation method, The parameters of BoB with segmentation method and BoHB with PCA Hashing strategy are strictly according to the configuration in [1] and [47], respectively.

**Evaluation Metric:** The performance of retrieval methods is evaluated by the normalized Discounted Cumulative Gain (nDCG) [102]. The definition of nDCG is

$$\text{nDCG}@k = \frac{1}{Z} \sum_{j=1}^k \frac{2^{\text{rel}(j)} - 1}{\log(1 + j)}, \quad (3.10)$$

where  $Z$  is used to normalize the calculated score, and  $\text{rel}(\cdot)$  evaluates the similarity between the query image and retrieval image in the repository.

**Retrieval Performance:** To tune the optimal value for  $\alpha$  in our attribute refinement model, we try several groups of parameters by using some evaluation samples as queries and calculate the performance of our retrieval system. Generally, the performance of our retrieval system comes to maximum when the value of  $\alpha$  is around 1.0. Therefore, we simply set  $\alpha$  as 1 in our experiment.

Figure 3.8 illustrates the result of baselines and our retrieval model with different input data. Generally, our retrieval method outperforms the baselines under most of the configurations. Specifically, two aspects can be observed. If we fix the data source of input as shown in Figure 3.8 (a), Figure 3.8 (b), and Figure 3.8 (c), our query-specific attribute refinement retrieval system is superior than the other two searching strategies, especially the retrieval strategy with the similarity of SIFT feature. On the other hand, if we fix the searching strategy, *i.e.*, the lines of the same color across Figure 3.8 (a), Figure 3.8 (b), and Figure 3.8 (c), the method with our detected result can achieve higher nDCG than the methods with the result of detection baselines. This observation implicitly indicates that our detection method is more accurate than the detection baselines

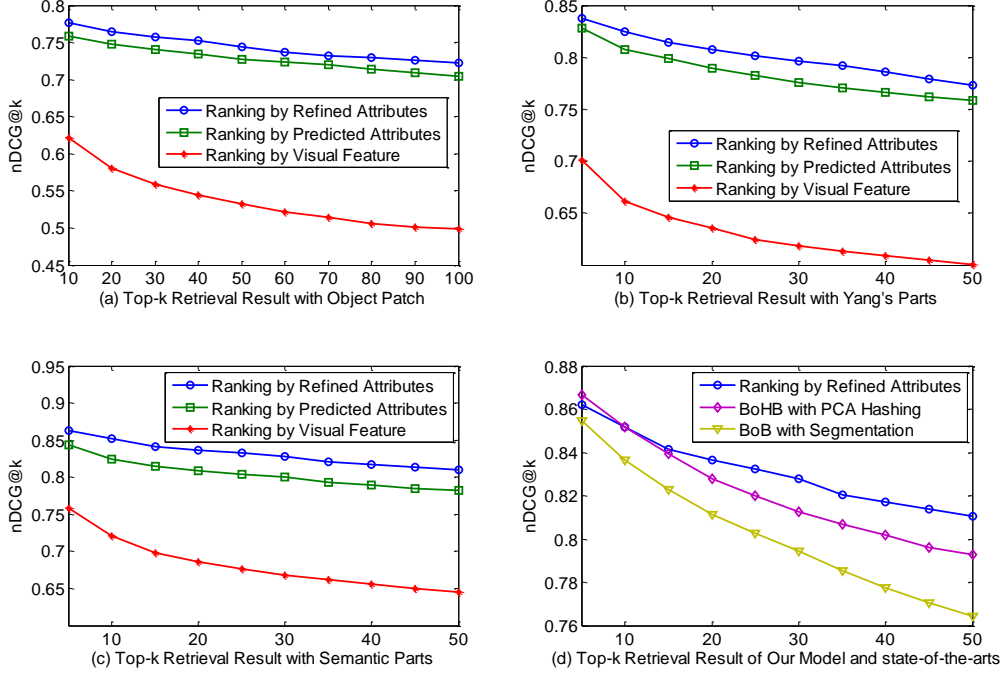


Figure 3.8: The nDCG@k of baselines and our proposed retrieval system, *i.e.*, query-specific attribute refinement retrieval method. The experiment results present that nDCG@k of specific retrieval method gradually decreases as the number of retrieval images increases, which matches our expectation as the later retrieval images usually are more irrelevant to query image. When fixing the number of retrieval images, our retrieval method outperforms the baselines under most of the experiment configurations. Specifically, two aspects can be observed. If we use the same input data as shown in each sub-figure, our query-specific attribute refinement method is superior than the other two searching strategies. If we use the same retrieval strategy, *i.e.*, the lines of the same color across sub-figure (a) to sub-figure (c), the retrieval method with our detected result (sub-figure (c)) outperforms the methods combined with detection baselines (sub-figure (a) and sub-figure (b)). This observation implicitly indicates the better performance of our detection method. In sub-figure (d), we can observe that the our retrieval result is also comparable to the state-of-the-art retrieval systems.

both in terms of parts and attributes. In practice, by using the short length of attribute probability vector as feature, the retrieval time is ignorable when comparing with the time for detection. Generally, our retrieval system spends about  $2.5 \sim 3$  seconds to retrieval a query image in  $500 \times 500$  pixels resolution.

In Figure 3.8 (d), we compare our query-specific attribute refinement method with additional two baselines: BoB with segmentation method, and BoHB with PCA Hashing ( $r=2$ ) and bounding reranking. Generally, the performance of our retrieval system is comparable to the two baselines. However, the performance of BoHB slightly outperforms our attribute refinement model when the number of retrieval images is less than 20. When the number of retrieval images is larger than 20, our attribute refinement model achieves better result. Basically, these two baselines use the low-level features, e.g. SURF or HoG, combining with boundary information by using different strategies. This fusion strategy may greatly contribute to the improvement of retrieval performance.

However, we claim that our retrieval system has some advantages comparing with BoHB and BoB. By using attribute related feature, our retrieval result is more explicit and expressive. Users can clearly observe the ranking criterion of our retrieval system. Moreover, our model can handle the non-rigid object retrieval problem, which can not be properly solved in BoHB and BoB. Meanwhile, our system is more tolerant of the discrepancy caused by viewpoint and rotation, which is not fully solved in BoHB with PCA Hashing and BoB with Segmentation.

**User Study:** To qualitatively evaluate our retrieval system, we conduct a user study on the demo systems to compare the retrieval result of our model and the baselines. Generally, 24 users of different careers are hired to score the result of different retrieval methods. Every query image and top-10 retrieval images returned by retrieval methods are presented as one group. Each user is required to view 50 groups of retrieval result, and answer the following questions by scoring each group from 0 to 10. Then, the average scores are calculated on the 50 groups of samples, and the mean score of 24 users with standard deviation is used to evaluate the performance of retrieval systems. The specific questions for this user study include:

Table 3.3: The average score and standard deviation of user study on the baselines and our retrieval system

Retrieval Model		Mean Score (Standard Deviation)			
Data Source	Retrieval Strategy	Q1	Q2	Q3	Q4
Object Patch	Rank by feat.	$5.9 \pm 0.54$	$4.3 \pm 0.78$	$4.1 \pm 1.13$	$7.4 \pm 0.43$
	Rank by attr.	$7.1 \pm 0.38$	$5.5 \pm 0.69$	$4.2 \pm 0.88$	$7.6 \pm 0.32$
	Refined attr.	$7.2 \pm 0.35$	$5.4 \pm 0.57$	$4.2 \pm 1.00$	$7.8 \pm 0.46$
Part-based Model	Rank by feat.	$6.5 \pm 0.48$	$6.2 \pm 0.75$	$5.9 \pm 0.96$	$6.7 \pm 0.44$
	Rank by attr.	$7.7 \pm 0.32$	$7.3 \pm 0.61$	$6.8 \pm 1.09$	$6.9 \pm 0.36$
	Refined attr.	$7.9 \pm 0.27$	$7.4 \pm 0.58$	$7.1 \pm 1.13$	$6.8 \pm 0.36$
Our Model	Rank by feat.	$7.5 \pm 0.41$	$7.6 \pm 0.70$	$7.2 \pm 1.04$	$6.5 \pm 0.48$
	Rank by attr.	$8.2 \pm 0.28$	$8.3 \pm 0.62$	$7.5 \pm 0.95$	$6.7 \pm 0.37$
	Refined attr.	$8.3 \pm 0.25$	$8.8 \pm 0.61$	$7.6 \pm 0.81$	$6.9 \pm 0.54$
BoB with Segmentation		$8.2 \pm 0.48$	$8.2 \pm 0.39$	$7.3 \pm 0.76$	$6.9 \pm 0.37$
BoHB with PCA Hashing		$8.1 \pm 0.41$	$8.4 \pm 0.55$	$7.3 \pm 0.60$	$7.2 \pm 0.44$

- Question 1: Is the retrieval result relevant to the query image in terms of attributes? Please score them according to heel shape, heel height, decoration, sole shape, tie style, topline, toe shape, vamp style, upper style, and closed mode (1 score for each attribute, 10 scores in total).
- Question 2: Does the retrieval result preserve the general style of query image? (10-8 scores: Fully Preserve. 7-5 scores: Partially Preserve. 4-3 scores: Slightly Preserve. 2-0 scores: Does not Preserve)
- Question 3: Does the retrieval result meet your performance expectation? (10-9 scores: Exceed. 8-7 scores: Meet. 6-5 scores: Partially Meet. 4-3 scores: Moderate. 2-0 scores: Does not Meet)
- Question 4: What do you think of the response time of retrieval system? (10-9 scores: Very Quick. 8-7 scores: Quick. 6-5 scores: Acceptable. 4-3 scores: Need Improvement. 2-0 scores: Unbearable)

The mean score and standard deviation of retrieval systems are presented in Table 3.3. Obviously, the question 1 and question 4 are more objective, while the question 2 and question 3 are more subjective, which can be observed from the standard deviation of user scores.

Generally, our retrieval system achieves higher score than most of baselines in terms of quality. The lower standard deviation of our system may also indicate

the stability of our retrieval result. Specifically, question 1 implicitly represents the accuracy of attribute prediction. It is interesting to point out that the scores of those retrieval systems coincide with the accuracy of attribute prediction in Table 3.2. This observation further reveals the efficacy of our detection model. Compared with BoB and BoHB, we claim that our detector can extract more distinctive feature of query images. In question 2, most users consider that our retrieval result can better preserve the general style of query images than most of the baselines. This may represent the operability of attribute-based retrieval strategy. Moreover, it should be noticed that our retrieval system achieves much more scores than the BoB and BoHB in this question. However, the overall score of question 3 is relatively low, which may be due to the high response time of our system. This is also represented in the question 4.

### 3.6 Chapter Summary

In this chapter, we proposed the Circle & Search retrieval system to search the semantically similar product instances in the repository by circling the interested product in daily photos as query. Taking footwear as example, our system contains two components, *i.e.*, the footwear detection and the footwear retrieval component. In detection component, by using the consistency between the footwear parts and semantic attributes, the detector can estimate the positions of parts and the values of attributes simultaneously. During the retrieval, the correlations between attributes are analysed and used to refine the predicted attribute values. Then, the refined attribute values are used to rank all the footwear instances in a query-specific way by computing the attribute distances. In the experiment, a large-scale footwear dataset is collected, and the experiment result on this dataset well demonstrated the effectiveness of our Circle & Search system. Compared with other retrieval systems, the retrieval result of our system is more expressive due to the semantic meaning of retrieval feature. Besides, the retrieval problem of non-rigid object can also be solved by our system due to the part-based property. Last but not least, our system is more tolerant of the discrepancy caused by changeable viewpoints and rotation. By defining the

tree-structures and part-attribute relations, we claim that our system can be extended to other domains, given the assumption that the attributes of domain object impact the visual appearance of parts.

In the next chapters, we propose to use the fine-grained attributes for the alignment of retrieval features learned by CNN, and apply this framework for the specific people description and clothing retrieval.



## Chapter 4

# Describing People by Clothing Attributes

In this chapter, we address the problem of describing people identity based on fine-grained clothing attributes. This is an important problem for many practical applications, such as identifying target suspects or finding missing people based on detailed clothing descriptions in surveillance videos or consumer photos. We approach this problem by first *mining* clothing images with *fine-grained* attribute labels from online shopping stores. A large-scale dataset is built with about one million images and fine-grained attribute sub-categories, such as various shades of color (e.g., watermelon red, rosy red, purplish red), clothing types (e.g., down jacket, denim jacket), and patterns (e.g., thin horizontal stripes, houndstooth). As these images are taken in ideal pose/lighting/background conditions, it is unreliable to directly use them as training data for attribute prediction in the domain unconstrained images captured, for example, by mobile phones or surveillance cameras. In order to bridge this gap, we propose a novel double-path deep domain adaptation network to model the data from the two domains jointly. Several alignment cost layers placed in-between the two columns ensure the consistency of the two domain features and the feasibility to predict unseen attribute categories in one of the domains. Finally, to achieve a working system with automatic human body alignment, we trained an enhanced R-CNN based detector to localize human bodies in images. Our extensive experimental evaluation demonstrates the effectiveness of the proposed approach for describing

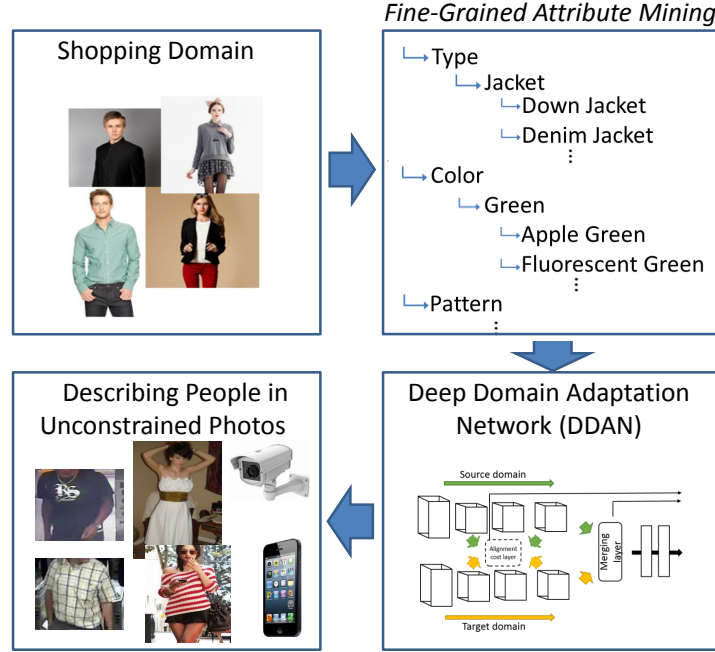


Figure 4.1: Overview of the proposed approach. We propose a novel deep domain adaptation method to bridge the gap between images crawled from online shopping stores and unconstrained photos. Another unique aspect of our work is the ability to describe people based on their fine-grained clothing attributes.

people based on fine-grained clothing attributes.

## 4.1 Introduction

Describing people *in detail* is an important task for many applications. For instance, criminal investigation processes often involve searching for suspects based on detailed descriptions provided by eyewitnesses or compiled from images captured by surveillance cameras. The FBI list of nationwide wanted bank robbers <sup>1</sup> has clear examples of such *fine-grained descriptions*, including attributes covering detailed color information (e.g., “light blue” “khaki”, “burgundy”), a variety of clothing types (e.g., “leather jacket”, “polo-style shirt”, “zip-up wind-breaker”) and also detailed clothing patterns (e.g., “narrow horizontal stripes”, “LA printed text”, “checkered”).

Traditional computer vision methods for describing people, however, have only focused on a small set of coarse-grained attributes. As an example, the recent work of Zhang et al. [124] achieves impressive attribute prediction per-

<sup>1</sup><https://bankrobbers.fbi.gov/>

formance in unconstrained scenarios, but only considers nine human attributes. Existing systems for fashion analysis [8, 121] and people search in surveillance videos [33, 109] also rely on a relatively small set of clothing attributes.

Our work addresses the problem of describing people with very fine-grained clothing attributes. In particular, we consider attribute sub-categories that differ in subtle details, including many shades of clothing color (e.g., “Apple Green”, “Fluorescent Green”, “Light Green”), different types of a particular garment (e.g., “Denim Jacket”, “Down Jacket”), and specific clothing patterns (e.g., “thin horizontal stripes”, “other types of stripes”). As far as we know, this is the first work to address this problem in a real scenario.

Directly tackling this problem is challenging because a large amount of annotated data is required to train such a large number of attribute models. In recent years, large-scale datasets such as ImageNet [16] and labelled Faces in the Wild [52] have been built by leveraging vast amounts of visual data available on the web. However, most of the images obtained from online sources are either unlabelled or weakly labelled, often requiring costly manual annotation. In this work, we draw attention to e-Commerce websites, such as *Amazon.com* and *TMALL.com*, which contain *structured descriptions* of products and can be considered a reliable source of annotation. By leveraging this rich source of data from online shopping stores, we are able to collect a large-scale annotated dataset with around one million images along with fine-detailed attribute sub-categories.

Some of the typical images from these online shops are shown in Figure 4.1. As can be seen, there is a large discrepancy between these samples and the samples from our application domain, i.e., unconstrained photos captured by, for example, surveillance cameras or mobile phones. The online shopping images are often depicted with ideal lighting, standard pose, high resolution, and good quality, whereas these conditions cannot be guaranteed for images captured in the wild. Thus we investigate whether it is possible to perform domain adaptation to bridge the gap between these two domains.

We look into the newest weapon of computer vision research – deep learning approaches, which have been applied very effectively for visual recognition problems e.g., the large scale visual recognition ImageNet challenge [65, 16], and the

object classification and detection tasks for PASCAL VOC datasets [39, 116]. Very recently, several works in computer vision have shown that it is generally effective to transfer a deep learning model learned from a large-scale corpus, e.g., ImageNet, to other tasks by using the activation maps of certain layers of Deep Convolutional Neural Networks (e.g., the second fully connected layer, FC2) [96, 51]. The underlying assumption of these methods is that the parameters of the low-level and mid-level network layers can be re-used across domains. As it may not be true for our domain adaptation problem, we aim to learn the domain-invariant hierarchical features directly, while transferring the domain information within intermediate layers. To this end, we design a specific double-path deep convolutional neural network for the domain adaptation problem. Each path receives one domain images as the input and they are connected through several alignment cost layers. These cost layers ensure that (1) the feature learning parameters for the two domains are not too far away and (2) similar labels encode similar high-level features.

The **contributions** of this work can be summarized as follows:

1. We target fine-grained attribute learning on a large-scale setting. Previous works only deal with a relatively small set of coarse-grained person attributes.
2. We collected a large-scale annotated dataset of garments, which contains around one million images and hundreds of attributes. As far as we know, this is the largest dataset for clothing analytics and attribute learning. We believe many applications can benefit from this dataset.
3. To bridge the gap between the two clothing domains considered in our work, we propose a specific double-path deep neural network which models the two domains with separate paths. Several additional alignment layers have been placed connecting the two paths to ensure the consistency of the two domain classifiers.
4. Our work is part of an actual product for people search in surveillance videos based on fine-grained clothing attributes.

## 4.2 Related Work

*Semantic visual attributes* have received significant attention by the computer vision community in the past few years [70, 28, 92, 68]. Among other applications, attributes have been used for zero-shot classification [70], visual search [63, 102], fine-grained categorization [6], and sentence generation from images [67]. Most of these methods rely on costly manual annotation of labels for training the attribute models. Notable exceptions include techniques that mine attributes from web data [10], including sources such as Wikipedia [98] and online books [17]. Our work follows this direction, but we focus on cross-domain attribute mining, where the data is mined from online shopping stores and then adapted to unconstrained environments, using a novel deep domain adaptation approach.

**Attribute Datasets.** There are only a few attribute datasets with fine-grained annotations, for example, datasets related to detailed descriptions of birds [111] and aircrafts [110]. We push this envelope by proposing a new dataset of fine-grained clothing attributes. Compared to other clothing datasets for fashion analysis [8, 121], our proposed dataset has a much larger set of garments, including attribute *sub-categories* and a massive volume of training images per class.

**Describing People by Attributes.** Predicting human attributes [68, 5, 125, 33] is important for many surveillance applications, such as person re-identification across cameras [71], suspect search based on eyewitness testimonies [33, 109], and identification based on soft-biometrics [56]. Our approach deals with a *fine-grained* set of clothing attributes, which is around 10x larger than most previous methods, and requires minimal manual labeling for attribute learning.

Extracting clothing attributes for *analysis of fashion images* is another topic that has recently attracted interest [8, 121, 62, 80]. Previous methods developed for this application domain often focus on the clothing segmentation problem, considering pictures depicted in relatively simple poses, against relatively clean backgrounds. In our work, we study the domain adaptation problem from “clean” clothing images obtained from online shopping stores to images captured

in unconstrained environments. Liu et al. [80] addressed a similar cross-domain clothing retrieval problem, but their work relies on a different methodology than ours, deals with a different application, and only considers a small set of coarse-grained attributes which are manually labelled.

**Deep Learning.** Deep Convolutional Neural Networks have recently achieved dramatic accuracy improvements in image classification [65], object detection [39], and many other computer vision areas, including attribute modeling [124, 85]. Recent improvements on deep learning include the use of drop-out [49] for preventing overfitting, more effective non-linear activation functions such as rectified linear units [40] or max-out [43], and richer modeling through Network-in-Network (NIN) [74]. In our work, we customize R-CNN and NIN for body detection, and propose a new deep domain adaptation approach to bridge the gap between the source and target clothing domain distributions.

**Domain Adaptation.** Many methods have been proposed for domain adaptation in visual recognition [42, 45, 41]. Recently, addressing this problem with deep neural networks has gained increased attention. The majority of existing approaches for domain adaptation or transfer learning with deep architectures rely on re-training the last few layers of the network using samples from the target domain, or instead performing *fine-tuning* of all layers using backpropagation at a lower learning rate [90, 96, 51]. However, these methods usually require a relatively large amount of training samples from the target domain to produce good results. In contrast, our method learns domain-invariant hierarchical features directly and transfers the domain information within intermediate layers, which we show to be much more effective. The work of Nguyen et al. [88] shares some of our motivations, but uses a different methodology based on dictionary learning.

A distinct method that was recently proposed for deep adaptation is DLID [11] which learns multiple unsupervised deep models directly on the source, target, and combined datasets, and uses a representation which is the concatenation of the outputs of each model as its adaptation approach. While this was shown to be an interesting approach, it is limited by its use of unsupervised deep structures, which have been unable to achieve the performance of supervised deep

CNNs. Our method instead uses a supervised double path CNN with shared layers. It is able to leverage the extensive labelled data available in the source domain using a supervised model without requiring a significant amount of labelled target data.

## 4.3 Dataset Preparation

Although there are a few existing fashion datasets in the research community [121, 79, 81], they are designed for the tasks of clothing parsing or human segmentation and no annotation of fine-grained attributes is included. Here we introduce our two sets of data and their statistics: (1) online shop dataset obtained by crawling large amount of annotated images from online shopping stores and (2) “street” dataset which consists of both web street images and sample videos from surveillance cameras.

### 4.3.1 Online Shop Dataset

#### Automatic Data Collection

We crawled a large amount of garment images from several large online shopping stores, e.g., Amazon.com and TMALL.com. We also downloaded the webpages which contain the images. These webpages can be parsed into  $\langle key, value \rangle$  pairs where each *key* corresponds to an attribute category, for example, “color” and the *value* specifies the attribute label, for example, “purplish red”. The total number of clothing images is 1,108,013 and it includes 25 different kinds of *keys*, i.e., attribute categories (e.g. type, color, pattern, season, occasion). The attribute labels are very fine-detailed. For instance, we can find more than ten thousand different *values* for the “color” category.

#### Data Curation

In consumer photos or images from surveillance cameras, it might be difficult or impossible for a person to differentiate some attributes that could otherwise be discriminated in the online shopping domain. For instance, a security guard would likely not be able to tell the difference between a suspect wearing a “ginger

yellow” shirt and another wearing a “turmeric” shirt. Therefore, we focus on a subset of the dataset mined from online shopping stores. We consider upper clothing only and select three attribute categories: type, color, and pattern. Several attributes are merged based on human perception. We also removed some attributes that are not well-defined, e.g., clothing images with “abstract patterns”. Finally, we removed attributes for which the number of collected images is less than 1000. As a result, we focus on a subset of the data containing 341,021 images and 67 attributes, including 15 kinds of clothing types, 31 colors and 21 kinds of patterns. We denote this dataset as *Online-data*.

### Building a Fine-grained Attribute Hierarchy

As described in section 4.3.1, the set of fine-grained attribute categories and labels mined from online shopping websites are given as a list of  $\langle key, value \rangle$  pairs without a hierarchical structure. We therefore organize this data into a semantic taxonomy of clothing attributes. We consider “type”, “color”, and “pattern” as the three higher-level categories. Each attribute is then classified into these three categories and further divided into semantic sub-categories. As an example, “wedding dress” and “sleeveless tank dress” are both sub-categories of “dress”, which is in turn a sub-category of “type”.

#### 4.3.2 Offline Street Dataset

Our unconstrained photo dataset, i.e., street domain dataset, consists of both web street images and videos from surveillance cameras.

**Web street images.** This data consists of two parts: (a) the standard Fashionate dataset [121], which consists of 685 street fashion images. This dataset has semantic segmentation labels (e.g., bags, pants), but no fine-grained clothing attributes. We fully annotated this dataset with our fine-grained attributes for evaluation purposes. We denote this data as *Street-data-a*. (b) the Parsing dataset [25] which consists of 8000 street fashion images. This dataset also has detailed segmentation labels, but we only used its images. We denote this data as *Street-data-b*.



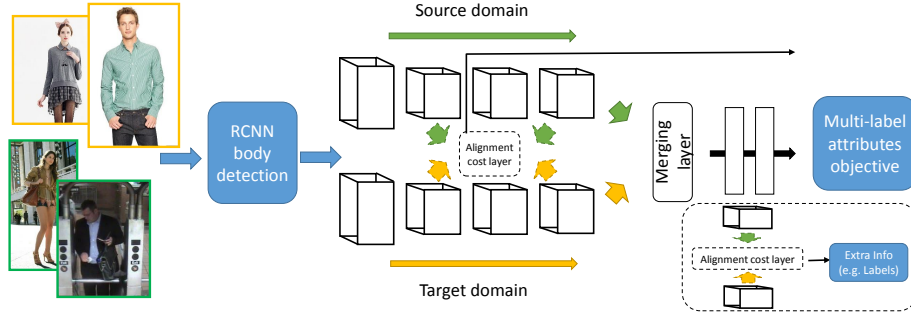


Figure 4.2: Overview of our proposed approach.

**Surveillance videos.** we consider 14 surveillance videos captured from two public train stations. The duration of each video is 10 minutes. These videos have different camera angles (e.g., captured from the train station platform, gateway, lobby). We manually annotated the bounding boxes of each person in the videos, using a step size of 60 frames. Thus the total number of frames/images we annotated is  $14 \times 10 \times 30 \times 60 / 60 = 4200$ . There are 6.2 people on average for each frame with reasonable size. We also annotated 120 frames with our fine-grained clothing attributes, using a region-of-interest where pedestrians have higher-resolution. These 120 frames were used as evaluation. The rest of the data (with bounding box annotation) was used as extra data for unsupervised training. We denote this dataset as *Street-data-c*.

It is worth noting that the “Street datasets” are relatively small considering the fine-grained attribute learning problem. It is impractical to directly learn the attributes from the street domain. The partially labelled street dataset will be fed into our learning framework as supervised training samples and evaluation ground-truth. These unlabelled data will be used as guiding samples to induce the network to fit the target domain feature distribution. More details will be discussed in the next section.

## 4.4 Approach

We now introduce our solution to tackle the problem of describing people based on fine-grained attributes, as shown in Figure 4.2. First, we introduce an improved version of the R-CNN body detector which effectively localizes the cloth-

ing area. We then describe our proposed approach for attribute modeling and domain adaptation.

#### 4.4.1 RCNN for Body Detection

Our body detection module is based on the R-CNN framework [39], with several enhancements made specifically for the clothing detection problem. It consists of three sub-modules. First, selective search is adopted to generate candidate region proposals. Then, a Network-in-Network (NIN) model is used to extract features for each candidate region. Finally, linear support vector regression (SVR) is used to predict the Intersection-over-Union (IoU) overlap of candidate patches with ground-truth bounding boxes. Next we introduce these components, and elaborate on the details of our enhancements.

**Region Proposals.** Due to the non-rigid property of clothing, standard selective search based on super-pixel proposal generation is shown to be more suitable to our detection task. Usually, about 2000 region proposals are generated per image. However, after analyzing the detection result by feeding all the region proposals, we find that lots of the false alarms are caused by the inappropriate region proposals, such as regions with small size or extreme aspect ratio. We tackle this problem by filter out the noisy proposals in the pre-processing step. According to the clothing size in the training images, we discard the candidate regions with inappropriate size and aspect ratio. After that, about 100 hypotheses are left, which considerably reduces the number of noisy proposal regions and thus accelerates the feature extraction procedure. Surprisingly, there are even about 4% increasement on detection accuracy after applying this filter strategy. However, in the R-CNN framework, the detection performance still largely depends on the performance of proposal generation method. In our future work, we may adopt the faster R-CNN [99] framework for our body detection module.

**Feature Extraction.** The NIN model [74] is used to extract the high-level features from the candidate regions. Briefly, this model is pre-trained on the Imagenet Challenge dataset (ILSVRC-2014 classification task), and then fine-tuned using a subset of clothing images from our data.

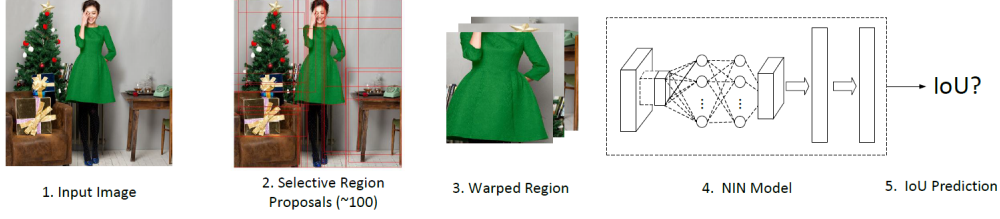


Figure 4.3: Enhanced R-CNN detection pipeline.

**Region IoU Prediction.** In the R-CNN framework, the positive samples in training are the candidate regions with relatively large IoUs overlapped with ground-truth objects. We claim that there are two shortcomings in this strategy. First, users are required to select a good IoU overlap threshold, which is crucial to the detection performance. Second, all image regions whose IoUs do not meet the threshold are discarded. However, we suggest that those regions are useful for the detection task. In our approach, instead of predicting a yes/no value for a given region, we actually predict its IoU overlap value. We used a linear regression model (SVR) in our implementation for predicting the region IoU using the features extracted by the fine-tuned NIN model.

In our implementation, we discretize the IoU values into ten intervals with a step of 0.1, and sample the equivalent training regions for each interval to balance the data during the training procedure. Lastly, the bounding box regression is employed to refine the selected proposal regions with the activation of the NIN fully-connected layer (FC2) as features.

#### 4.4.2 Deep Domain Adaptation

Although we have collected a large scale dataset with fine-grained attributes, these images are taken in ideal pose/lighting/background conditions, so it is unreliable to directly use them as training data for attribute prediction in the domain of unconstrained images captured, for example, by mobile phones or surveillance cameras. In order to bridge this gap, we design a specific double-path deep convolutional neural network for the domain adaptation problem. Each path receives one domain image as the input, i.e., the street domain and the shop domain images. Each path consists of several convolutional layers which are stacked layer-by-layer and normally higher layers represent higher-level con-

cept abstractions. Both of the two network paths share the same architecture, e.g., the same number of convolutional filters and number of middle layers. This way, the output of the same middle layer of the two paths can be directly compared. We further connect these paths through several alignment cost layers where the cost function is correlated with the similarity of the two input images. These alignment cost layers are included to ensure that (1) the feature learning parameters for the two domains are not too far away and (2) the high-level features have sufficient similarity along with the label consistency.

We also design a merging layer whose input is from the two network paths, which are merged and share parameters in the subsequent layers. This design is used to deploy the model after the co-training. We take the merging operation as the simple *max* operation, i.e.  $f(X_s, X_t) = \max(X_s, X_t)$ . So we can simply drop out this layer at testing time.

### Alignment Layer Cost Function

We present the alignment layer cost function in the following form:

$$f(s, t) = \|X_s - X_t\| \times \lambda \phi(s, t), \quad (4.1)$$

where  $X_i = w_i \otimes y_i$  is the activation from the connection layer, e.g., the convolutional layer or the fully connected layer and  $\phi(s, t)$  is a correlation function. We can directly obtain the gradient of this cost w.r.t. the connection layer to reduce computational cost. If we consider a fully supervised domain adaptation problem, we can set the correlation function as the label similarity, e.g.  $\phi(l_s, l_t) = \exp\{-\frac{\|l_s - l_t\|^2}{\gamma}\}$ , where  $l_s$  and  $l_t$  are the attribute label vectors for the source and target domain images, respectively. If we consider a semi-supervised or unsupervised learning problem, we can assume this function is defined by additional prior information, e.g., the visual similarity function. Note that we work on multiple attribute categories at the same time, i.e. we model the attribute classifiers simultaneously. The final overall learning objective of the DDAN is defined as a combination of a multi-label classification objective and multiple alignment regularization terms.

## Discussion

It is worth noting the the following unique properties of the proposed DDAN: Consider a simplified CNN-based classification function, i.e.,  $y = f(g(x), w)$  where  $w$  are the classifier parameters (e.g., the final logistic regression layer) and  $g(x)$  is the deep feature generator. In our domain adaptation problem, DDAN tries to align the target domain feature generator  $g_{tgt}(x)$  with the source domain feature generator  $g_{src}(x)$ . As opposed to traditional domain adaptation approaches which try to align the features by finding a suitable subspace [42, 34], DDAN aims to align the high/middle level features directly during the training step of feature learning.

**Comparison with deep learning fine-tuning framework:** The popular fine-tuning framework usually takes the output of the last layer of the network as a feature and performs additional training for the new tasks or performs fine-tuning on the whole original network without dropping the original objective function. The former case is not suitable for our problem as we don't have enough diverse training samples to re-train the target domain classifier. The latter case is equivalent to adapting the classifier  $y = f(g(x), w)$  to  $y_{tgt} = f_{tgt}(g_{tgt}(x), w_{tgt})$ . The proposed DDAN has two distinct properties over this solution: (1) it puts an additional regularization term on the adaptation process which seeks the feature agreement w.r.t. the prior information, e.g., the label consistency or visual similarity. (2) the learned feature generator  $g_{tgt}$  has consistent output with the source domain feature so that we can directly apply new attribute classifiers to unseen labels learned from the source domain without additional cost.

**Comparison with Siamese network:** The structure of the proposed DDAN is similar to the Siamese network, often used for verification tasks. Both networks have two paths and a merging step. But the functions of these two key modules are quite different. The Siamese network calculates the difference of the two input channels and there is no back propagation channel to constrain the middle level representation. Moreover, the weights of two sub-network in Siamese network are shared, while the weights in our proposed DDAN are separated for modelling the data of two domains.

**The role of the alignment cost layers:** Our main motivation is that instead of learning the domain adaptation at the classifier level, we aim to learn the domain invariant feature directly through the hierarchical feature learning model. The alignment cost layers connecting the two paths at a higher level plays a fundamental role in this process. The unshared lower level features model each domain specific data distribution while we constrain the higher level features to be consistent w.r.t. the label/prior information.

## 4.5 Experiments

In this section, we present our implementation details, provide an extensive experimental analysis for evaluating our proposed approach, and showcase a couple of applications. We will be referring to the datasets described in section 4.3: *Online-data*, *Street-data-a*, *Street-data-b*, and *Street-data-c*.

### 4.5.1 Implementation Details

#### Network configuration

For each path of the DDAN, we configure the network with the same setting of the standard AlexNet [65], i.e. 5 convolutional layers and 3 fully connected layers, and the same filter numbers and neuron sizes. We have put the alignment cost layer at Conv5, FC1 and put the merging layer at FC2. During testing, we simply drop out the merging layer and use the target domain network. For the retrieval task, we used the FC2 output for both source and target domains.

We used a modified code of **cuda-convnet**<sup>2</sup>. We modified the code to enable the support of multi-attribute co-training, i.e. we trained the three attribute categories at the same time. We also modified the code to support the correlation function.

**Initialization:** We tried out two initialization methods: (1) using random Gaussian values and (2) using the model learned from the ImageNet dataset as the initialization for both source and target domains. Generally, the second option gave us more stable results and fast convergence. Thus, we used this

---

<sup>2</sup>[code.google.com/p/cuda-convnet/](http://code.google.com/p/cuda-convnet/)

initialization method in the next sections. Normally 20 epochs are enough for the network training.

### Learning setting

**Supervised training.** If we have the annotation from the target domain, we define the correlation function  $\phi$  for the alignment cost layers as the similarity at the label level. For each source domain image, we first select a set of target domain images with the closest label distance. We then further rank this set according to the visual similarity w.r.t the source image, and select the first one as the assignment. This pair is fed into the network as the input of the two paths. We set the network label as the source domain (shopping) as it has a much larger and diverse amount of data.

**Unsupervised training.** If we don't have the annotation from the target domain, we use prior knowledge to define the similarity between the source and target images. In our experiments, we used a large amount of unannotated street images. In practice, we perform an iterative procedure for the training. At each epoch, we use the linear product similarity of the current FC2 layer features to find the nearest neighbour of a given source domain sample in the target domain dataset. The correlation function  $\phi$  is also defined as this linear similarity. Then the source domain image and its neighbour are fed into the network. After one epoch, we re-calculate the similarity using the updated model. This procedure iterates until the preset epoch number is reached.

#### 4.5.2 Exp1: RCNN body detection

For our body detection experiment, we annotated 4,000 images of the *Online-data* and 2,000 images of the *Street-data-b* with clothing bounding boxes for training. The *Street-data-a* dataset was used for validation. We compare the result of our enhanced RCNN-NIN detector with two baselines: (1) Deformable Part-based Model(DPM) [30] and (2) RCNN using a traditional CNN model [65] pre-trained on ImageNet [16] as feature generator. All of the baselines and the proposed method are tested using the same dataset.

As the performance of detection greatly affects the attribute learning, we set

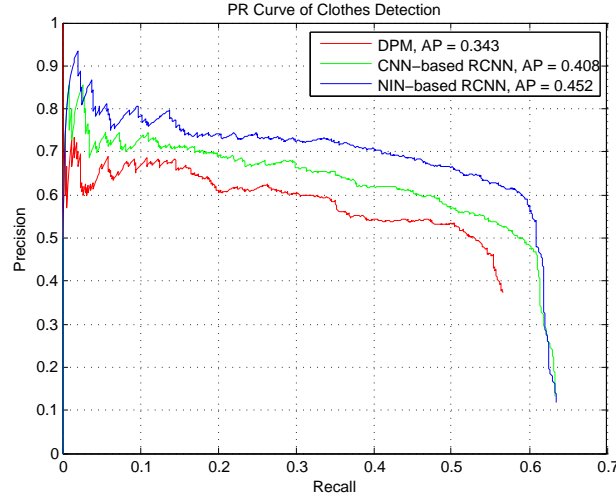


Figure 4.4: Precision-recall curves for body detection results on *Street-data-a*.

Table 4.1: Fine-grained attributes classification results for *Street-data-a*.

Street-data-a	CNN	CNN-FC2	CNN-FT	DDAN-S	DDAN-U
Type-T1	25.44 %	22.12 %	32.53 %	31.02 %	<b>33.42 %</b>
Type-T1-b	29.68 %	25.67 %	37.9 %	36.20 %	<b>38.92 %</b>
Color-T1	16.23 %	10.02 %	22.87 %	25.21 %	<b>27.39 %</b>
Color-T1-b	20.18 %	14.43 %	27.21 %	30.46 %	<b>32.30 %</b>
Pattern-T1	73.11 %	60.91 %	<b>76.2 %</b>	75.31 %	74.13 %
Pattern-T1-b	73.39 %	63.20 %	<b>76.6 %</b>	76.01 %	74.90 %

a strict evaluation metric for body detection. More specifically, we consider a detection to be correct only if the overlap of the prediction and ground-truth bounding boxes is over 0.6 instead of 0.5 as common in standard evaluation. We evaluate the performance of our body detector on *Street-data-a* with Precision Recall curves as shown in Figure 4.4. We also report the Average Precision (AP) result. As can be seen, our RCNN-NIN detector consistently outperforms the baselines (RCNN-NIN AP 0.452 vs DPM AP 0.343 and conventional-RCNN AP 0.408 ).

#### 4.5.3 Exp2: DDAN for Fine-grained Attribute Classification

We consider the following methods for comparison: (1) **CNN**, where we directly apply the model learned from the source domain to the target domain. (2) **CNN-FC2**, where we use the FC2 layer of a CNN model [65] as the features for training a classifier using street domain images. (3) **CNN-FT**, i.e., CNN fine-tuning, which keeps the original shop domain objective function, and then feeds



Table 4.2: Fine-grained attributes classification results for *Street-data-c*.

Street-data-c	CNN	DDAN-U
Type-T1	41.53 %	<b>48.32 %</b>
Type-T1-b	48.31 %	<b>55.12 %</b>
Color-T1	5.08 %	<b>15.34 %</b>
Color-T1-b	5.93 %	<b>18.87 %</b>
Pattern-T1	70.34 %	<b>72.45 %</b>
Pattern-T1-b	71.19 %	<b>75.90 %</b>

the network with the street domain training samples. (4) **DDAN-S**, where we use the supervised setting of the proposed DDAN. (5) **DDAN-U**, where we use the unsupervised setting of the DDAN, without using any annotated target domain data. It is worth noting that we didn’t use any tricks which are commonly used in the ImageNet challenges (e.g., multiple models ensembles, data augmentation, etc.) to improve the performance.

Regarding the *Street-data-a* dataset, we split it into two halves and used the first half as the target domain training samples, and the other half for testing. We used *Street-data-b* as extra data during training. Regarding the *Street-data-c* dataset, we tested only the unsupervised setting of DDAN, as we have very limited fine-grained attribute annotation data for this dataset.

**Evaluation metrics:** We used Top-1 (T1) and Top-1-base (T1-b) accuracy as the evaluation metrics, defined as follows. Top-1 accuracy is the standard evaluation metric for general classification problems. As claimed in Sec 4.3, we are working with a fine-grained attribute list. The attributes themselves naturally fall into a hierarchical structure. If the prediction and the groundtruth share the same immediate father node in the hierarchy, we consider it as a correct prediction. In this case, the accuracy we get is Top-1-base accuracy.

**Result analysis:** We present our results in Tables 4.2 and 4.1. For the *Street-data-c*, i.e., the surveillance video dataset, we only report the results of the fully unsupervised setting due to the lack of annotation for this domain. We can see consistent improvement of **DDAN-U** over directly applying the source domain CNN model with big margin. It shows that the feature learning can benefit from large amounts of unannotated target domain data. For the *Street-data-a* result, i.e., the street photo dataset result, we can see that our domain adaptation methods outperform the baselines **CNN**, **CNN-FC2** and

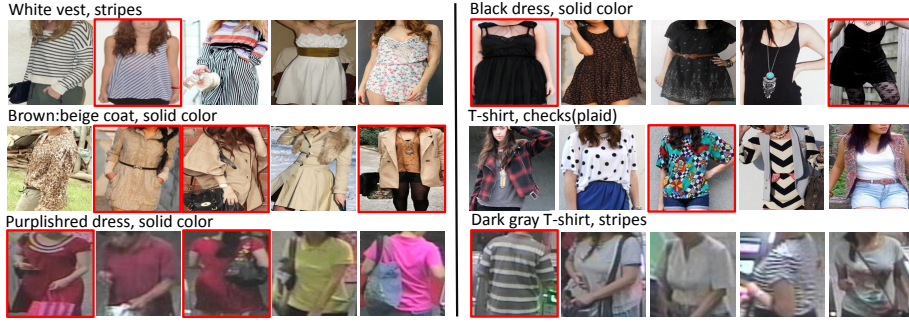


Figure 4.5: Application 1: Attribute-based people search. We rank the images according to their attribute scores. The top-5 ranking results for each query are exhibited. Top 2 rows results are from *Street-data-a*, and the bottom results are from *Street-data-c*. The images that exactly match the query are marked with red bounding box. Best viewed in original pdf file.



Figure 4.6: Application 2: Street2Shop clothing retrieval. Top 2 rows results are from *Street-data-a*, and the bottom two rows are from *Street-data-c*. We output the top 3 retrieval results for both datasets. Best viewed in original pdf file.

**CNN-FT.** We can see that **DDAN-U** achieves the best results on most of the categories.

Overall, we notice that we can achieve much better results for “Type” and “Pattern” than “Color” categories, especially in the surveillance scenario. It is reasonable as “Color” is very sensitive w.r.t. the lighting condition in the wild, while “Type” and “Pattern” are more related to the “Shape” of the garments. Our domain adaptation framework reduces the gap between the two domains.

#### 4.5.4 Application 1: Attribute-based People Search

Here we showcase a few examples of attribute-based people search using the proposed system in Figure 4.5, e.g. finding people wearing a black-stripes T-shirt. We rank images based on the sum of the attribute confidence scores. We

only show the top-5 ranked images due to space limitation. The images that exactly match the query are marked with red bounding box.

#### 4.5.5 Application 2: Street2Shop Clothing Retrieval

As discussed in Sec 4.4.2, one major advantage of the proposed DDAN is that the output features share the same distribution of the source domain. So we can directly calculate the similarity of the two domain images without finding the common feature subspace or metric space. It provides great convenience for clothing retrieval – we can easily find the most similar online shopping clothes by looking at the feature similarity, e.g. the linear product distance. Some exemplar results are shown in Figure 4.6. We showcase the results for both *Street-data-a* and *Street-data-c* datasets. We output the top 3 retrieval results for both datasets.

## 4.6 Chapter Summary

In this chapter, we presented a novel deep domain adaptation network for the problem of describing people based on fine-grained clothing attributes. To handle the problem of domain discrepancy, the double patch deep network is used to model the data from two domain jointly, which achieves better performance than the conventional fine-tuning pipeline. According to the problem setting, this network can be applied to supervised or unsupervised problem by defining different alignment cost layers. As far as we know, this is the first work to address the problem of cross-domain people description in a real scenario. Our experiments show the advantage of the proposed approach over the baselines. We also showcased practical applications of this work, including the people retrieval via text query and Street2Shop clothing retrieval. The qualitative results of those examples indicate the effectiveness of our DDAN on practical applications.

In the following chapter, we present a similar framework for cross-domain retrieval feature learning. Instead of using attributes as similarity criterion, we extract the semantic representation of images from the fine-grained attributes. Based on the semantic representation, the cross-domain retrieval features are

learned via a learning-to-rank framework.

## Chapter 5

# Cross-domain Attribute-aware Image Retrieval

In this chapter, we address the problem of cross-domain image retrieval, considering the following practical application: given a user photo depicting a clothing image, our goal is to retrieve the same or attribute-similar clothing items from online shopping stores. This is a challenging problem due to the large discrepancy between online shopping images, usually taken in ideal lighting/pose/background conditions, and user photos captured in uncontrolled conditions. To address this problem, we propose a Dual Attribute-aware Ranking Network (DARN) for retrieval feature learning. More specifically, DARN consists of two sub-networks, one for each domain, whose retrieval feature representations are driven by semantic attribute learning. We show that this attribute-guided learning is a key factor for retrieval accuracy improvement. In addition, to further align with the nature of the retrieval problem, we impose a triplet visual similarity constraint for learning to rank across the two sub-networks. To train our network, we collect a large-scale dataset with fine-grained attributes. Specifically, we exploit customer review websites to crawl a large set of online shopping images and corresponding offline user photos with fine-grained clothing attributes, i.e., around 450,000 online shopping images and about 90,000 exact offline counterpart images of those online ones. All these images are collected from real-world consumer websites reflecting the diversity of the data modality, which makes this dataset unique and rare in the academic community. We extensively evaluate

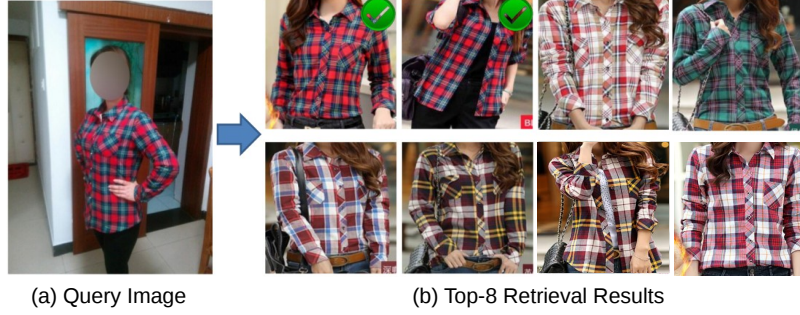


Figure 5.1: Cross-domain clothing retrieval. (a) Query image from daily photos. (b) Top-8 product retrieval results from the online shop domain. The proposed system finds the exact match clothing (first two images) and ranks the ones with similar attributes as top results.

the retrieval performance of networks in different configurations. The top-20 retrieval accuracy is **doubled** when using the proposed DARN other than the current popular solution using pre-trained CNN features only (0.570 vs. 0.268).

## 5.1 Introduction

There is a long history of methods for content-based image retrieval in the field of computer vision [36, 3]. However, little work has been devoted to the problem of cross-domain image retrieval, defined as follows: given a query image from domain  $X$ , retrieve similar images from a database of images belonging to domain  $Y$ .

This problem setting arises in many important applications. For example, mobile product image search [100] aims at identifying a product, or retrieving similar products from the online shop domain based on a photo captured in unconstrained scenarios by a mobile phone camera. In surveillance applications, a security guard may be interested in retrieving images of a suspect from a specific camera given a query image from another camera.

In this chapter, we address the problem of cross-domain product retrieval by taking clothing products as a concrete use case. Given an offline clothing image from the “street” domain, our goal is to retrieve the same or similar clothing items from a large-scale gallery of professional online shop images, as illustrated in Figure 5.1.

Due to the huge impact for e-commerce applications, there is a growing interest in methods for clothing retrieval [60, 86, 80, 114] and outfit recommendation [55]. The majority of these methods, however, do not model the discrepancy between the user photos and clothing images from online shopping stores. Though metric learning methods [44, 66] can be used for domain adaptation, their performance largely depends on the existing features. Another barrier also occurs because of the lack of large annotated training sets containing user photos and desired retrieved images from online shopping websites.

In order to tackle the training data issue, we observe that there is a large number of customer review websites, where people post their pictures wearing the clothing they have purchased. Therefore, it is possible to crawl the offline clothing images uploaded by the users with the links to the online shop product images. As a result, we created a dataset containing tens of thousands of online-offline clothing image pairs obtained from the user review pages. These image pairs are very rare in both academic and industry as they reveal the real discrepancy of images across scenarios. In addition, we have also obtained corresponding fine-grained clothing attributes (e.g., clothing color, collar pattern, sleeve shape, sleeve length, etc.) from the available online product description, without significant annotation cost. As data pre-processing, in order to remove the impact of cluttered backgrounds, which predominantly exist for the offline images, we employ an enhanced R-CNN detector to localize the clothing area in the image, with some refinements particularly made for the clothing detection problem.

For addressing the problem of cross-domain retrieval, we propose a novel Dual Attribute-aware Ranking Network (DARN) for retrieval feature learning. DARN consists of two sub-networks with similar structure. Each of the two domain images are fed into each of the two sub-networks. This specific design aims to diminish the discrepancy of online and offline images.

The two sub-networks are designed to be driven by semantic attribute learning, so we call them attribute-aware networks. The intuition is to create a powerful semantic representation of clothing in each domain, by leveraging the vast amounts of data annotated with fine-grained clothing attributes. Tree-structure

layers are embedded into each sub-network for the comprehensive integration of attributes and their full relations. Specifically, the low-level layers of the sub-network are shared for learning the low-level representation. Then, a set of fully connected layers in a tree-structure are used to construct the high-level component, with each branch modelling one attribute.

Based on the learned semantic features from each attribute-aware network, we incorporate the learning-to-rank objective to further enhance the retrieval feature representation. Specifically, the triplet ranking loss is used to constrain the feature similarity of triplets, i.e., the feature distance between the online-offline image pair must be smaller than that of offline image and any other dissimilar online images.

Generally, the retrieval features from DARN have several advantages compared with the deep features of other works [59, 19]. (1) By using the dual-structure network, our model can handle the cross-domain problem more appropriately. (2) In each sub-network, the scenario-specific semantic representation of clothing is elaborately captured by leveraging the tree-structure layers. (3) Based on the semantic representation, the visual similarity constraint enables more effective feature learning for the retrieval problem.

In summary, the main **contributions** of this work are:

1. We collect a unique dataset composed of cross-scenario image pairs with fine-grained attributes. The number of online images is about 450,000, with additional 90,000 offline counterparts collected. Each image has about 5-9 semantic attribute categories, with more than a hundred possible attribute values. This online-offline image pair dataset provides a training/testing platform for many real-world applications related to clothing analytics. We are planning to release the full dataset to the community for research purposes only.
2. We propose the Dual Attribute-Aware Ranking Network which simultaneously integrates the attributes and visual similarity constraint into the retrieval feature learning. We design tree-structure layers to comprehensively capture the attribute information and their full relations, which pro-



vides a new insight on multi-label learning. We also introduce the triplet loss function which perfectly fits into the deep network training.

3. We conduct extensive experiments proving the effectiveness and robustness of the framework and each one of its components for the clothing retrieval problem. The top-20 retrieval accuracy is **doubled** when using the proposed DARN other than using pre-trained CNN feature only (0.570 vs. 0.268). The proposed method is general and could be applied to other cross-domain image retrieval problems.

## 5.2 Related Work

**Fashion Datasets.** Recently, several datasets containing a wide variety of clothing images captured from fashion websites have been carefully annotated with attribute labels [121, 22, 82, 55]. These datasets are primarily designed for training and evaluation of clothing parsing and attribute estimation algorithms. In contrast, our data is comprised of a large set of clothing image pairs depicting user photos and corresponding garments from online shopping websites, in addition to fine-grained attributes. Notably, this real-world data is essential to bridge the gap between the two domains.

**Visual Analysis of Clothing.** Many methods [77, 37, 103, 76, 75] have been recently proposed for automated analysis of clothing images, spanning a wide range of application domains. In particular, clothing recognition has been used for context-aided people identification [37], fashion style recognition [62], occupation recognition [103], and social tribe prediction [69]. Clothing parsing methods, which produce semantic labels for each pixel in the input image, have received significant attention in the past few years [121, 22]. In the surveillance domain, matching clothing images across cameras is a fundamental task for the well-known person re-identification problem [71, 101].

Recently, there is a growing interest in the methods for clothing retrieval [60, 86, 80, 114] and outfit recommendation [55]. Most of those methods do not model the discrepancy between the user photos and online clothing images. An exception is the work of Liu et al. [80], which follows a very different methodol-

ogy than ours based on part-based alignment and features derived from sparse reconstruction, and does not exploit the richness of our data obtained by mining images from customer reviews.

**Visual Attributes.** Research on attribute-based visual representations have received renewed attention by the computer vision community in the past few years [70, 28, 93, 110]. Attributes are usually referred as semantic properties of objects or scenes that are shared across categories. Among other applications, attributes have been used for zero-shot learning [70], image ranking and retrieval [102, 63, 53], fine-grained categorization [6], scene understanding [94], and sentence generation from images [67].

Related to our application domain, Kovashka et al. [63] developed a system called “WhittleSearch”, which is able to answer queries such as “Show me shoe images like these, but sportier”. They used the concept of relative attributes proposed by Parikh and Grauman [93] for relevance feedback. Attributes for clothing have been explored in several recent papers [8, 9, 5]. They allow users to search visual content based on fine-grained descriptions, such as a “blue striped polo-style shirt”.

Attribute-based representations have also shown compelling results for matching images of people across domains [101, 72]. The work by Donahue and Grauman [18] demonstrates that richer supervision conveying annotator rationales based on visual attributes, can be considered as a form of privileged information [108]. Along this direction, in our work, we show that cross-domain image retrieval can benefit from feature learning that simultaneously optimizes a loss function that takes into account visual similarity and attribute classification.

**Deep Learning.** Deep convolutional neural networks have achieved dramatic accuracy improvements in many areas of computer vision [65, 39, 105]. The work of Zhang et al. [124] combined poselet classifiers [5] with convolutional nets to achieve compelling results in human attribute prediction. Sun et al. [105] discovered that attributes can be implicitly encoded in high-level features of networks for identity discrimination. In our work, we instead explicitly use attribute prediction as a regularizer in deep networks for cross-domain image retrieval.

Attribute categories	Examples (total number)
Clothes Button	Double Breasted, Pullover, ... (12)
Clothes Category	T-shirt, Skirt, Leather Coat ... (20)
Clothes Color	Black, White, Red, Blue ... (56)
Clothes Length	Regular, Long, Short ... (6)
Clothes Pattern	Pure, Stripe, Lattice, Dot ... (27)
Clothes Shape	Slim, Straight, Cloak, Loose ... (10)
Collar Shape	Round, Lapel, V-Neck ... (25)
Sleeve Length	Long, Three-quarter, Sleeveless ... (7)
Sleeve Shape	Puff, Raglan, Petal, Pile ... (16)

Table 5.1: Clothing attribute categories and example values. The number in brackets is the total number of values for each category.



Figure 5.2: Some examples of online-offline image pairs, containing images of different human pose, illumination, and varying background. Particularly, the offline images contain many selfies with high occlusion.

Existing approaches for image retrieval based on deep learning have outperformed previous methods based on other image representations [3]. However, they are not designed to handle the problem of cross-domain image retrieval. Several domain adaptation methods based on deep learning have been recently proposed [51, 11]. Related to our work, Chen et al. [9] used a double-path network with alignment cost layers for attribute prediction. In contrast, our work addresses the problem of cross-domain retrieval feature learning, proposing a novel network architecture that learns effective features for measuring visual similarity across domains. We note that other domain adaptation methods [44, 66] could even be applied on top of our learned features to further refine retrieval results.

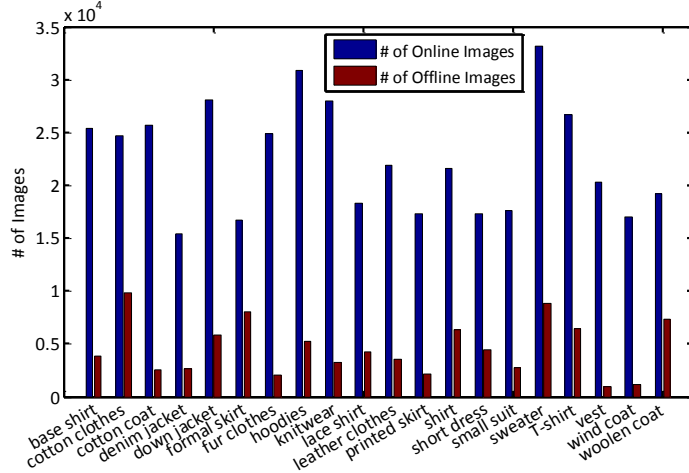


Figure 5.3: The distribution of online-offline image pairs.

### 5.3 Data Collection

We have collected about 453,983 online upper-clothing images in high-resolution (about  $800 \times 500$  on average) from several online-shopping websites. Generally, each image contains a single frontal-view person. From the surrounding text of images, semantic attributes (e.g., clothing color, collar shape, sleeve shape, clothing style) are extracted and parsed into  $\langle key, value \rangle$  pairs, where each *key* corresponds to an attribute category (e.g., color), and the *value* is the attribute label (e.g., red, black, white, etc). Then, we manually pruned the noisy labels, merged similar labels based on human perception, and removed those with a small number of samples. After that, 9 categories of clothing attributes are extracted, and the total number of attribute values is 179. As an example, there are 56 values for the color attribute.

The specified attribute categories and example attribute values are presented in Table 5.1. This large-scale dataset annotated with fine-grained clothing attributes is used to learn a powerful semantic representation of clothing, as we will describe in the next section.

Recall that the goal of our retrieval problem is to find the online shopping images that correspond to a given query photo in the “street” domain uploaded by the user. To analyze the discrepancy between the images in the shopping scenario (online images) and street scenario (offline images), we collect a large set of offline images with their online counterparts. The key insight to collect

this dataset is that there are many customer review websites where users post photos of the clothing they have purchased. As the link to the corresponding clothing images from the shopping store is available, it is possible to collect a large set of online-offline image pairs.

We initially crawled 381,975 online-offline image pairs of different categories from the customer review pages. Then, after a data curation process, where several annotators helped removing unsuitable images, the data was reduced to 91,390 image pairs. For each of these pairs, fine-grained clothing attributes were extracted from the online image descriptions. Some examples of cropped online-offline image pairs are presented in Figure 5.2. As can be seen, each pair of images depict the same clothing, but in different scenarios, exhibiting variations in pose, lighting, and background clutter. The distribution of the collected online-offline images is illustrated in Figure 5.3. Generally, the number of images of different categories in both scenarios are almost in the same order of magnitude, which is helpful for training the retrieval model.

In summary, our dataset is suitable to the clothing retrieval problem for several reasons. First, the large amount of images enables effective training of retrieval models, especially deep neural network models. Second, the information about fine-grained clothing attributes allows learning of semantic representations of clothing. Last but not least, the online-offline images pairs bridge the gap between the shopping scenario and the street scenario, providing rich information for real-world applications.

## 5.4 Approach

The unique dataset introduced in the previous section serves as the fuel to power up our attribute-driven feature learning approach for cross-domain retrieval. Next we describe the main components of our proposed approach, and how they are assembled to create a real-world cross-domain clothing retrieval system.

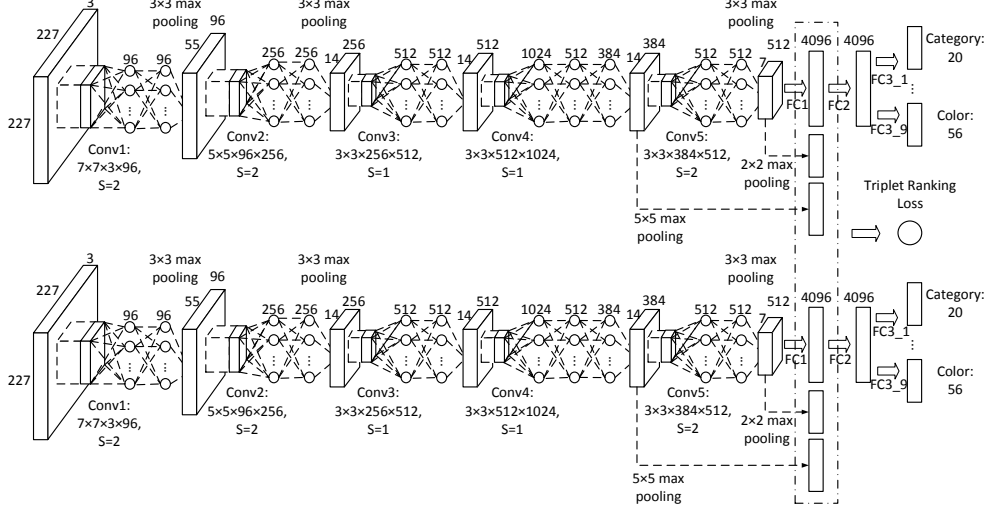


Figure 5.4: The specific structure of DARN, which consists of two sub-networks for images of the shopping scenario and street scenario, respectively. In each sub-network, it contains a NIN network, including all the convolutional layers, followed by two fully connected layers. The tree-structure layers are put on top of each network for attribute learning. The output features of each sub-network, i.e., FC1, Conv4-5, are concatenated and fed into the triplet ranking loss across the two sub-networks.

#### 5.4.1 Dual Attribute-aware Ranking Network

In this section, the Dual Attribute-aware Ranking Network (DARN) is introduced for retrieval feature learning. Compared to existing deep features [59, 19], DARN simultaneously integrates semantic attributes with visual similarity constraints into the feature learning stage, while at the same time modeling the discrepancy between domains.

##### Network Structure

The structure of DARN is illustrated in Figure 5.4. Two sub-networks with similar Network-in-Network (NIN) structure [74] construct the foundation of DARN. During training, the images from the online shopping domain are fed into one sub-network, and the images from the street domain are fed into the other. Each sub-network aims to represent the domain-specific information and generate high level comparable features as output. The NIN model in each sub-network consists of five stacked convolutional layers followed by MLPCConv (Multi-layer Perceptron Convolutional) layers as defined in [74], and two fully

connected layers (FC1, FC2). To increase the representation capability of the intermediate layer, the fourth layer, named Conv4, is followed by two MLPConv layers.

On top of each sub-network, we add tree-structured fully-connected layers to encode information about semantic attributes. Given the semantic features learned by the two sub-networks, we further impose a triplet-based ranking loss function, which separates the dissimilar images with a fixed margin under the framework of *learning to rank*. The details of semantic information embedding and the ranking loss are introduced next.

### Semantic Information Embedding

In the clothing domain, attributes often refer to the specific description of certain parts (e.g., collar shape, sleeve length) or clothing (e.g., clothes color, clothes style). Complementary to the visual appearance, this information can be used to form a powerful semantic representation for the clothing retrieval problem. To represent the clothing in a semantic level, we design tree-structure layers to comprehensively capture the information of attributes and their full relations.

Specifically, we transmit the FC2 response of each sub-network to several branches, where each branch represents a fully-connected network to model each attribute separately. In this tree-structured network, the visual features from the low-level layers are shared among attributes; while the semantic features from the high-level layers are learned separately. The neuron number in the output-layer of each branch equals to the number of corresponding attribute values (see Table 5.1). Since each attribute has a single value, the cross-entropy loss is used in each branch. Note that the values of some attributes may be missing for some clothing images. In this case, the gradients from the corresponding branches are simply set to zero.

During the training stage, the low-level representation of clothing images is extracted layer by layer. As the activation transfers to the higher layers, the representation becomes more and more abstract. Finally, the distinctive characteristic of each attribute is modelled in each branch. In the back-propagation, the gradient of loss from each attribute w.r.t. the activation of FC2 layer are

summed up and transferred back for weight update.

### Learning to Rank with Semantic Representation

In addition to encoding the semantic representation, we apply the learning to rank framework on DARN for retrieval feature learning. Specifically, the triplet-based ranking loss is used to constrain the feature similarity of image triplets. Denoting  $a$  and  $b$  the features of an offline image and its corresponding online image respectively, the objective function of the triplet ranking loss is:

$$Loss(a, b, c) = \max(0, m + \text{dist}(a, b) - \text{dist}(a, c)), \quad (5.1)$$

where  $c$  is the feature of the dissimilar online image,  $\text{dist}(\cdot, \cdot)$  represents the feature distance, e.g., Euclidean distance, and  $m$  is the margin, which is empirically set as 0.3 according to the average feature distance of image pairs. Basically, this loss function imposes that the feature distance between an online-offline clothing pair should be less than that of the offline image and any other dissimilar online image by at least margin  $m$ .

In this way, we claim that the triplet ranking loss has two advantages. First and obviously, the desirable ranking ordering can be learned by this loss function. Second, as the features of online and offline images come from two different sub-networks, this loss function can be considered as the constraint to guarantee the comparability of features extracted from those two sub-networks, therefore bridging the gap between the two domains.

Similar to [19], we found that the response of FC1 layer, i.e., the first fully connected layer, achieves the best retrieval result. Therefore, the triplet ranking loss is connected to the FC1 layer for feature learning. However, the response from the FC1 layer encodes global features, implying that subtle local information may be lost, which is specially relevant for discriminating clothing images. To handle this problem, we claim that local features captured by convolutions should also be considered. Specifically, the max-pooling layer is used to down-sample the response of the convolutional layers into  $3 \times 3 \times f_n$ , where  $f_n$  is the number of filters in the  $n$ -th convolutional layer. Then, the down-sampled



response is vectorized and concatenated with the global features. Lastly, the triplet ranking loss is applied on the concatenated features of every triplet. In our implementation, we select the pooled response map of Conv4 and Conv5, i.e., the last two convolutional layers, as local features.

## Discussion

It is worth to discuss the connections and differences between our DDAN and DARN frameworks. Generally, both frameworks share similar structure, which consists of two sub-networks, and each sub-network learns the image representation from one domain. In feed-forward procedure, the low-level image features of specific domain are extracted layer by layer by their corresponding sub-network, and the alignment layer is added at the high-level component of network for domain adaptation. In the back-propagation, the gradient of loss w.r.t to domain-specific images are back-propagated into their corresponding sub-network for updating the weights of individual sub-network. These two strategies guarantee that the images representation from two sub-networks are discriminative yet comparable. Last but not least, both models can be applied on the task of attribute classification and image retrieval.

However, there are still some intrinsic differences between those two frameworks. First of all, the DDAN framework is proposed for cross-domain attribute prediction, and the DARN is proposed for the cross-domain image retrieval problem. In the DDAN framework, the alignment layer for domain adaptation can be integrated via a supervised manner or unsupervised manner, and the attributes are directly used as guidance to evaluate the similarity of two instances in a supervised manner. Another distinct characteristic of DDAN is that the data of target domain can be unlabelled. While, the semantic attributes in DARN framework are learned via the proposed tree-structure layers, where the individual information and joint relations of attributes are simultaneously modelled in a holistic semantic representation vector. One drawback is that this network requires that the target domain images are fully annotated. To learn the comparable features, the triplet ranking loss is applied on the semantic representation of two domains images. In this way, the features from alignment layer in DARN framework,

i.e., FC1 layer, cannot only learn the ranking order of images, but also integrate semantic information. During the training procedure, the DDAN only models the relations between two similar/dissimilar images, while the DARN learns the relations among a triplet.

#### 5.4.2 Clothing Detection

As a pre-processing step, the clothing detection component aims to eliminate the impact of cluttered backgrounds by cropping the foreground clothing from images, before feeding them into DARN. Our method is an enhanced version of the R-CNN approach [39], which has recently achieved state-of-the-art results in object detection.

Analogous to the R-CNN framework, clothing proposals are generated by selective search [106], with some unsuitable candidates discarded by constraining the range of size and aspect ratio of the bounding boxes. Similar to Chen et al. [9], we process the region proposals by a NIN model. However, our model differs in the sense that we use the attribute-aware network with tree-structured layers as described in the previous section, in order to embed semantic information as extra knowledge. We show in our experiments that this model yields superior results.

Based on the attribute-aware deep features, support vector regression (SVR) is used to predict the intersection over union (IoU) of clothing proposals. In addition, strategies such as the discretization of IoU on training patches, data augmentation, and hard example mining, are used in our training process. As post-processing, bounding box regression is employed to refine the selected proposals with the same features used for detection.

#### 5.4.3 Cross-domain Clothing Retrieval

We now describe the implementation details of our complete system for cross-domain clothing retrieval.

**Training Stage.** The training data is comprised of online-offline clothing image pairs with fine-grained clothing attributes, as described in Section 5.3. The clothing area is extracted from all images using our clothing detector, and

then the cropped images are arranged into triplets.

In each triplet, the first two images are the online-offline pairs, with the third image randomly sampled from the online training pool. As the same clothing images have a unique ID, we sample the third online image until getting a different ID than the online-offline image pair. Several such triplets construct a training batch, and the images in each batch are sequentially fed into their corresponding sub-network according to their scenarios. We then calculate the gradients for each loss function (cross-entropy loss and triplet ranking loss) w.r.t. each sample, and empirically set the scale of gradients from those loss functions as 1. Lastly, the gradients are back propagated to each individual sub-network according to the sample domain.

We pre-trained our network as well as the baseline networks used in the experiments on the ImageNet dataset (ILSVRC-2014), as this yields improved retrieval results when compared to random initialization of parameters.

**End-to-end Clothing Retrieval.** We have set up an end-to-end real-time clothing retrieval demo on our local server. In our retrieval system, 200,000 online clothing images cropped by the clothing detector are used to construct our retrieval gallery. Given the cropped online images, the concatenated responses from FC1 layer, pooled Conv4 layer, and pooled Conv5 layer of one sub-network of DARN corresponding to shop scenario are used as the representation features. The same processes are operated on the query image, except that the other sub-network of DARN is used for retrieval feature extraction. We then  $l_2$  normalize the features from different layers for each image. The PCA is used to reduce the dimension of normalized features (17,920-D for DARN with Conv4-5) into 4,096-D, which conducts a fair comparison with other deep features using FC1 layer output only. Based on the pre-processed features, the Euclidean distance between query and gallery images is used to rank the images according to the relevance to the query.

In this system, one drawback is that the retrieval result depends on the performance of detector. Though a simple solution is to use the whole images for network training, we claim that this solution may only work for the online images, while the offline images are too cluttered to be directly fed into the

network for training. One example can be found in the query image of Fig 5.1, in which the clothing part only occupies partial region of the query image, and the lateral part of clothing is visible. Therefore, we claim that the alignment of foreground part in our retrieval system is necessary.

## 5.5 Experiments

### 5.5.1 Experimental Setting

**Dataset:** For training the clothing detector, 7,700 online-offline images are sampled from our dataset as positives and labelled with bounding boxes. The person-excluded images from the PASCAL VOC 2012 [27] detection task are used as negatives. Another 766 images are annotated to test the detectors.

For the retrieval experiment, about 230,000 online images and 65,000 offline images are sampled for network training. In the training process, each offline image and its online counterpart are collected, with the dissimilar online image randomly sampled from the 230,000 online pool to construct a triplet. Note that the third images in different epochs are shuffled to be different for the same online-offline pair. For testing, we used 1,717 online-offline image pairs. To make the retrieval result convincing, the rest 200,000 online images are used as the retrieval gallery.

**Baselines:** For clothing detection, we compare the performance of Deformable Part-based Model (**DPM**) [31] and different R-CNN versions with different models, including AlexNet (**Pre-trained CNN**) [65], **Pre-trained NIN**, and the Attribute-aware Network (**AN**). To evaluate the contribution of SVR, we compare the performance of SVR and SVM based on the AlexNet.

For clothing retrieval, the approach using Dense-SIFT (DSIFT) + fisher vector (FV) is selected as traditional baseline. Specifically, the bin size and stride for DSIFT are 8 and 4, respectively. The descriptor dimension is reduced to 64 by PCA. In the encoding step, two dictionaries with 64 and 128 centers are constructed, which lead to the 8,192 and 16,384 dimensions of FV representation.

To analyze the retrieval performance of deep features, we compare pre-trained networks including AlexNet (**pre-trained CNN**) and **pre-trained NIN**. We

Detection Model	Online AP	Offline AP	Top-20 Acc
DPM	0.049	0.017	0.297
Pre-trained CNN+SVM	0.520	0.412	0.560
Pre-trained CNN+SVR	0.545	0.452	0.567
Pre-trained NIN+SVR	0.601	0.477	0.588
AN+SVR	<b>0.744</b>	<b>0.683</b>	<b>0.635</b>

Table 5.2: AP of detection models on online-offline images and its corresponding top-20 retrieval accuracy on a subset of the data.

evaluate each individual component of our proposed approach. We denote our overall solution as Dual Attribute-aware Ranking Network (**DARN**), the solution without dual structure as Attribute-aware Ranking Network (**ARN**), the solution without dual structure and the ranking loss function as Attribute-aware Network (**AN**).

We further evaluate the effectiveness of DARN in terms of different configurations w.r.t. the features used, i.e., **DARN** using the features obtained from FC1, **DARN with Conv4** using the features from FC1+Conv4, and **DARN with Conv4-5** using the features from FC1+Conv4+Conv5. It is worth noting that the dimension of all features are reduced to 4096 by PCA to have a fair comparison.

**Evaluation Metrics:** We used two metrics to measure the performance of retrieval models. (1) the top-k retrieval accuracy in which we denote a hit if we find the exact same clothing in the top  $k$  results otherwise a miss, and (2) Normalized Discounted Cumulative Gain ( $NDCG@k$ ), considering  $NDCG@k = \frac{1}{Z} \sum_{j=1}^k \frac{2^{rel(j)} - 1}{\log(j+1)}$ , where  $rel(j)$  is the relevance score of the  $j^{th}$  ranked image, and  $Z$  is a normalization constant. The relevance score  $rel(j)$  of query image and  $j^{th}$  ranked image is the number of their matched attributes divided by the total number of query attributes.

### 5.5.2 Clothing Detection Improving Clothing Retrieval Performance

We used Average Precision (AP) to evaluate clothing detection. Since the detection performance is important to our network learning, a more strict IoU threshold, i.e., 0.7, is selected. The AP of detection results on online and offline

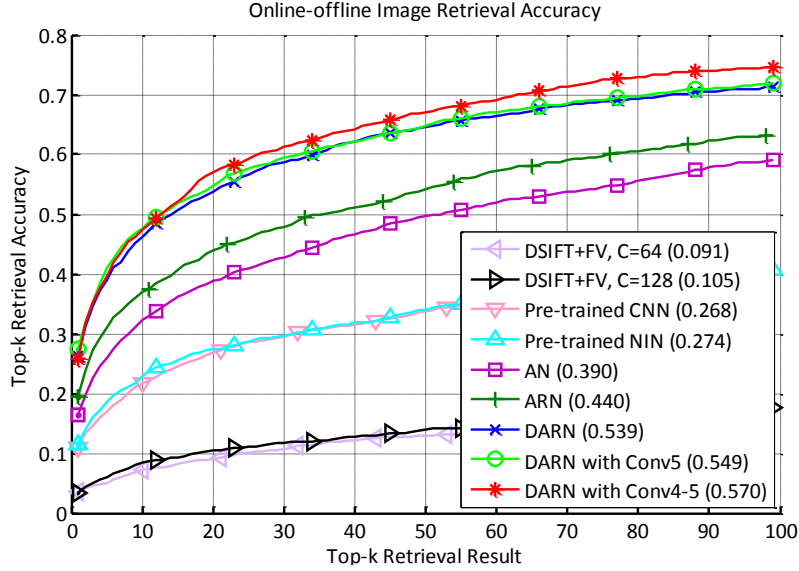


Figure 5.5: The top-k retrieval accuracy on 200,000 retrieval gallery. The number in the parentheses is the top-20 retrieval accuracy.

images is presented in Table 5.2. Generally, the performance of every detector on the online images is better than that on offline images, which indicates the complexity of offline images. We can observe that our proposed AN with SVR is superior than other baselines. DPM achieves the lowest AP, which may be due to less discriminative features and its incapability to handle clothing with huge distortion. By comparing the performance of CNN with SVM and SVR, we can find the effectiveness of SVR in the R-CNN framework. Furthermore, the detection performance is further improved by replacing the CNN with pre-trained NIN. Lastly, the AN with SVR achieves 74.4% and 68.3% AP on the online and offline images respectively, which is significantly better than the runner-up.

To evaluate the impact of various detectors on retrieval, we compare the top-20 retrieval accuracy of DARN with Conv4-5 by feeding different detection results. We sampled 10,000 online images from the full set as retrieval gallery for this test. The results are presented in Table 5.2. As can be seen, more precise detection leads to more accurate retrieval results.

### 5.5.3 Cross-domain Clothing Retrieval Evaluation

We give full detailed top-k retrieval accuracy results for different baselines as well as our proposed methods in Figure 5.5. We vary  $k$  as the tuning parameter

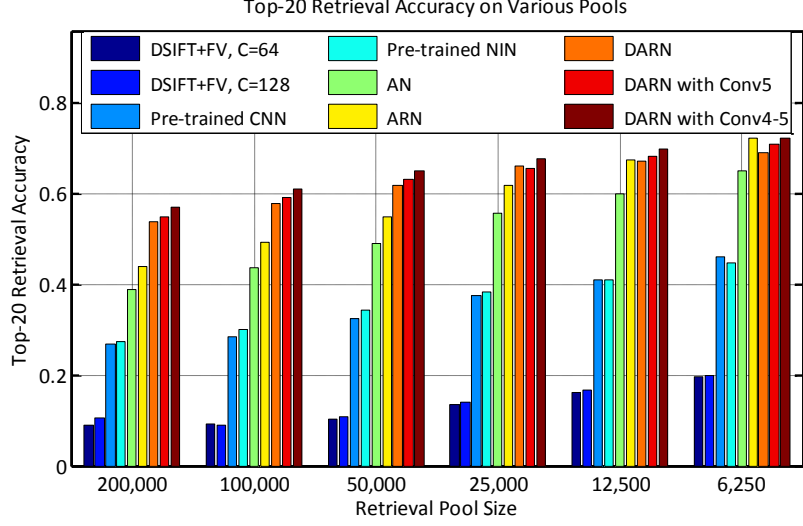


Figure 5.6: The top-20 retrieval accuracy on different sizes of retrieval galleries.

as it is an important indicator for a real system. We also list the top-20 retrieval accuracy of each model in the parentheses.

Compared to the baselines, we notice that all the deep features significantly outperform the traditional features, i.e., Dense-SIFT with FV encoding. For the deep features, the top-k accuracy of pre-trained NIN is slightly better than that of pre-trained CNN. Based on the pre-trained NIN, we evaluate the contributions of tree-structured layers, triplet ranking, and dual-structure.

Generally, the retrieval performance is gradually improved by applying the NIN structure, semantic information, learning to rank framework, and the dual-structure. The top-20 retrieval accuracy of AN increases 11.6% after fine-tuning on pre-trained NIN with attributes. This attests the effectiveness of attributes for image retrieval. By introducing the triplet ranking loss, the top-20 accuracy of ARN achieves another 5.0% increment.

Compared with a single model, the dual-structure network greatly improves the retrieval performance, i.e., the top-20 retrieval accuracy of DARN improves 9.9% when compared with ARN. The retrieval performance also slightly benefits from the local features, which can be observed by comparing the DARN and DARN with local features, i.e., DARN with Conv5 and DARN with Conv4-5. Some retrieval examples by DARN with Conv4-5 are illustrated in Figure 5.7.

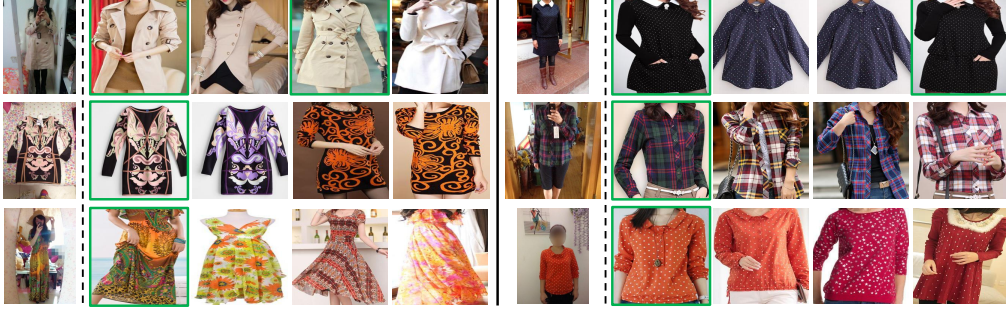


Figure 5.7: The top-4 retrieval result of DARN with Conv4-5. The images in first column are the queries, and the retrieved images with green boundary are the same clothing images. Best viewed in original pdf file.

#### 5.5.4 Attribute-aware Clothing Retrieval Evaluation

One key advantage of the proposed approach is the attribute-aware nature. The learned features have strong semantic meaning. Therefore, we should expect that the retrieval result should present strong attribute-level matching in terms of retrieval accuracy.

To evaluate this argument, we use NDCG@K to calculate the attribute-level matching. More specifically, we define the relevance score in NDCG as the attribute matching between the query and retrieval results divided by the total number of query attributes. We present the result in Table 5.3. Compared with traditional features, the retrieval result of deep features contains more similar attributes to the queries.

#### 5.5.5 Showing the Robustness: Performance vs. Retrieval Gallery Size

To further demonstrate the robustness our method, we show the top-20 retrieval accuracy of different retrieval models by tuning the retrieval gallery size in Figure 5.6.

We calculate the accuracy increment ratio of some representatives to evaluate the robustness of features. Intuitively, the smaller increase ratio indicates the better robustness of features. Specifically, the top-20 retrieval accuracy of traditional features, pre-trained NIN, ARN, and DARN increase by 115.4%, 63.8%, 64.5%, and 28.2% from largest retrieval gallery to smallest gallery, respectively.



Retrieval Model	NDCG@20
DSIFT + FV, C = 64	0.290
DSIFT + FV, C = 128	0.289
Pre-trained CNN	0.367
Pre-trained NIN	0.370
AN	0.415
ARN	0.442
DARN	0.494
DARN with Conv5	0.499
DARN with Conv4-5	0.505

Table 5.3: The NDCG@20 result evaluating the attribute level match on 200,000 retrieval gallery.

This observation demonstrates that the DARN can learn much more robust features than the baselines.

### 5.5.6 System Running Time

Our retrieval system runs on a server with the Intel i7-4930K CPU (@ 3.40GHz) with 12 cores and 65 GB RAM memory, with two GTX Titan GPU cards. On average, the attribute-aware ranking feature extraction process costs about 13 seconds per 1,000 images. Given a cropped query, it costs about 0.21 second for feature extraction and clothing retrieval in our retrieval experiment.

## 5.6 Chapter Summary

We have presented the Dual Attribute-aware Ranking Network for the problem of cross-domain image retrieval. Different from previous approaches, our method simultaneously embeds semantic attribute information and visual similarity constraints into the feature learning stage, while modeling the discrepancy of the two domains. We demonstrated our approach in a practical real-world clothing retrieval application, showing substantial improvement over other baselines. In addition, we created a unique large-scale clothing dataset which should be useful to many other applications.

## Chapter 6

# Conclusions and Future Works

This thesis explores the problem of cross-domain image retrieval by exploiting the semantic representation of images. Following the canonical pipeline of image retrieval, we highlight our contributions to the attribute-aware feature learning in three manners. In the traditional framework, the attributes are integrated as the high-order relations among parts for semantic object detection, and the predicted attribute probabilities vectors are used as the retrieval features. In the deep learning framework, we provide two methods for the integration of attributes in our proposed dual-structure network. One method is proposed to utilize the attributes to measure the instance similarity in the alignment cost layer, which bridges the representations of two analogous objects. In the other method, the tree-structure layers are designed for the simultaneous learning of multi-attributes, so that the responses of the joint layer can capture the comprehensive information of multi-attributes. Meanwhile, we collect a series of large-scale datasets with online-offline images and fine-grained attributes, which provide a meaningful platform for the evaluation of academic research works. In this chapter, we briefly summarize the main contents of this thesis, and discuss our future directions.

### 6.1 Thesis Conclusions

In Chapter 3, we introduce the attribute-aware part-based model for object detection and the query-specific attribute refinement model for retrieval features

learning. The main idea of the semantic object detection, on one hand, is to use the attributes as the high-order relations to maintain the consistency of appearance information among relevant parts. On the other hand, the precise localization of parts can enhance the prediction of semantic attributes. In details, we construct the part of our detection model as an “AND-OR” hierarchical mixture. Inside the mixture, the probability vector of attribute values belonging to a specific object is softly assigned to relevant parts, which constructs the first-level “AND” mixture. Within each component of the first-level mixture, the spatial distribution of parts is exclusively modelled by second-level mixture in the “OR” relation. Those parts are connected by a tree-structure model for the representation of objects, and the deformation information between parts is captured by the spring models. Due to the non-convex property of this model, the EM-based approach is used for optimization. Starting from the initialization, the localization of parts and the prediction of attributes are alternatively optimized by fixing one another. When the locations of parts are fixed, the attributes can be predicted via a multi-class classifier fed with the concatenation of part features. With the predicted attribute probability vectors, the localization of parts can be solved by the dynamic programming. Usually, those procedures can converge in 2-3 iterations. Based on the predicted attributes, the query-specific attribute refinement model is used to refine the attribute probability vector according to the co-occurrence of attributes in the retrieval gallery. By using the refined attribute probability vector as the retrieval feature, this framework achieves compelling results on the problem of cross-domain footwear retrieval.

In Chapter 4, we further push forward the application of fine-grained attributes on the problem of cross-domain people description. Rather than using attributes as retrieval features, we propose to align the representations of analogous objects by using attributes as the similarity metric, so that the similar objects have closer representations. Specifically, to eliminate the domain discrepancies, a double-path deep domain adaptation network is proposed to model the images of one domain in one sub-network. To ensure the consistency of two domain features, the alignment cost layer with customized cost functions is proposed and placed in-between the two columns of sub-networks. In the su-

pervised setting, the attributes are used to constrain the distance of analogous image pairs. For the semi-supervised problem, the cost function can be defined as the visual similarity of two images, and the network training can be implemented in an iterative manner, where the visual similarities of images are the distances of network responses in the previous iteration. In this way, the prediction of attribute categories in the source domain can be transferred to the target domain due to the comparability of image features. In the experiment, we extensively compare the performance of the pre-trained CNN model, the CNN model with the conventional fine-tuning manner, and our proposed DARN model. The experimental results indicate the superior performance of our framework for people description based on fine-grained attributes. To demonstrate the effectiveness of our framework in industry, we further apply this framework on two real-world applications, i.e., attribute-based people search and Street2Shop clothing retrieval.

In Chapter 5, we address the problem of cross-domain image retrieval by taking advantage of a dual-structure ranking network for retrieval feature learning. Analogous to DDAN, the dual-structure network is employed for the cross-domain learning. However, this framework is distinctively characterized by the attribute-aware property of retrieval features and the triplet ranking constraint. In details, the tree-structure layers are designed to simultaneously learn the multiple attributes of domain-specific images, and the semantic representation of images can be comprehensively captured by the network response. Based on the learned representation, the triplet ranking loss is used for retrieval feature learning. Different from the alignment cost layer, this loss function imposes that the feature distance between a similar pair should be less than that of a dissimilar pair by at least a pre-defined margin. By considering the similarity of dissimilar pairs, this loss function is proved to be more effective than the alignment cost layer. Lastly, to elaborately describe the object, we suggest that the response of high-level convolutional layers can perfectly compensate the insufficiency of fully connection layers in local feature representation.

Meanwhile, we considerably enhance the performance of a popular object detection framework, i.e., R-CNN. Basically, the NIN model is used to extract the

features for object detection, and the detection problem is modelled in a regression manner in our framework. In contrast to R-CNN, the NIN is fine-tuned on the clothing images with fine-grained attributes via our proposed tree-structure layers. Taking advantage of the attribute-aware property of the network, the detection performance is significantly improved. Note that this enhanced framework is also applied on the problem of people description in Chapter 4. In the experiment, we evaluate the contribution of each component, and conclude that those components gradually improve the retrieval performance on the cross-domain clothing retrieval.

## 6.2 Future Works

The fine-grained attributes are proved to be useful for the problem of cross-domain image retrieval. Though success has been achieved by attribute-driven learning, we suggest that there still exist several limitations:

- The efficiency of our retrieval systems is mainly limited by the speed of object detection. Though some strategies, e.g., proposal pruning, have been used for detection acceleration, the speed of object detection can still be enhanced in several aspects, e.g., proposal generation, multi-scale feature extraction.
- With the tree-structure layers, the attribute information can be comprehensively captured in our DARN model. However, one drawback is that the attributes are still attached to the full images of objects. Analogous to the attribute-aware part-based model, we claim that the retrieval performance may be further improved if the attribute-part relation can be applied.
- The problem of cross-domain image retrieval suffers from many impacts of domain discrepancies in terms of lighting, pose, viewpoints, which the still images cannot describe. To handle the problem of extreme situations, we suggest that the videos captured in unconstrained situations should be used for data collection.

Based on the aforementioned limitations, we propose the future directions for our image retrieval methods:

- Most recently, some efforts [38, 97] have been made for the problem of fast object detection. However, the efficiency of those methods largely depends on the proposal generation methods, e.g. selective search [107], or dense sliding windows. Instead of using object proposals, we suggest a much more efficient way that the object response maps can be directly generated by the original images. Though this method may degrade the detection accuracy, we claim that the semantic representation of objects can compensate for the performance degradation without losing efficiency.
- Inspired by the work of human parsing via deep learning [78, 73], we claim that the general objects can be decomposed into several semantic parts. Based on those parts, we suggest that the fine-grained attributes can be attached to some relevant parts, instead of the entire objects, for semantic representation learning. In this way, the representation of objects becomes more flexible and explainable. By extracting some discriminative parts and attributes, the problems caused by the viewpoint and rotation can be eliminated, and the retrieval systems with customized queries become more feasible, e.g., searching “a coat with black hoodie and short sleeve”.
- Different from still images, the videos provide a multi-view perspective and the spatio-temporal information for object description. However, those information are still under-exploitation in the field of image retrieval. To fully explore the potential of videos, we would like to employ the Recurrent Neural Network (RNN) [50] for retrieval feature learning. Owing to the property of attributes shared among objects, we suggest that the attributes can still be embedded into the RNN framework for domain adaptation.

# Bibliography

- [1] R. Arandjelovic and A. Zisserman. Smooth object retrieval using a bag of boundaries. In *International Conference on Computer Vision*, 2011.
- [2] R. Arandjelovic and A. Zisserman. All about vlad. In *Computer Vision and Pattern Recognition*, 2013.
- [3] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky. Neural codes for image retrieval. In *Proceedings of the European Conference on Computer Vision*, 2014.
- [4] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *Proceedings of the European Conference on Computer Vision*. 2010.
- [5] L. Bourdev, S. Maji, and J. Malik. Describing People: A Poselet-Based Approach to Attribute Classification. In *International Conference on Computer Vision*, 2011.
- [6] S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie. Visual Recognition with Humans in the Loop. In *Proceedings of the European Conference on Computer Vision*, 2010.
- [7] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.
- [8] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *Proceedings of the European Conference on Computer Vision*. 2012.
- [9] Q. Chen, J. Huang, R. Feris, L. Brown, J. Dong, and S. Yan. Deep domain adaptation for describing people based on fine-grained clothing attributes. In *Conference on Computer Vision and Pattern Recognition*, 2015.

- [10] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *International Conference on Computer Vision*, 2013.
- [11] S. Chopra, S. Balakrishnan, and R. Gopalan. Dlid: Deep learning for domain adaptation by interpolating between domains. In *Proceedings of International Conference on Machine Learning workshop on challenges in representation learning*, 2013.
- [12] C. Dagli, S. Rajaram, and T. Huang. Leveraging active learning for relevance feedback using an information theoretic diversity measure. *Image and Video Retrieval*, pages 123–132, 2006.
- [13] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Conference on Computer Vision and Pattern Recognition*, 2005.
- [14] N. Dalal and B. Triggs. INRIA person dataset. *Online: <http://pascal.inrialpes.fr/data/human>*, 2005.
- [15] F. de Natale and C. Dagli. Content-based image retrieval by feature adaptation and relevance feedback. . . . *IEEE Transactions on*, 2007.
- [16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li. ImageNet: A large-scale hierarchical image database . In *Conference on Computer Vision and Pattern Recognition*, 2009.
- [17] S. K. Divvala, A. Farhadi, and C. Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *Conference on Computer Vision and Pattern Recognition*, 2014.
- [18] J. Donahue and K. Grauman. Annotator rationales for visual recognition. In *International Conference on Computer Vision*, 2011.
- [19] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint [arXiv:1310.1531](https://arxiv.org/abs/1310.1531)*, 2013.
- [20] J. Dong, Q. Chen, J. Feng, K. Jia, Z. Huang, and S. Yan. Looking inside category: subcategory-aware object recognition. *Transactions on Circuits and Systems for Video Technology*, 2014.
- [21] J. Dong, Q. Chen, Z. Huang, J. Yang, and S. Yan. Parsing based on parselets: A unified deformable mixture model for human parsing. *Transactions on Pattern Analysis and Machine Intelligence*, 2015.



- [22] J. Dong, Q. Chen, X. Shen, J. Yang, and S. Yan. Towards unified human parsing and pose estimation. In *Conference on Computer Vision and Pattern Recognition*, 2014.
- [23] J. Dong, Q. Chen, W. Xia, Z. Huang, and S. Yan. A deformable mixture parsing model with parselets. In *International Conference on Computer Vision*, 2013.
- [24] J. Dong, Q. Chen, S. Yan, and A. Yuille. Towards unified object detection and segmentation. In *European Conference on Computer Vision*. 2014.
- [25] J. Dong, Q. Chen, S. Yan, and A. Yuille. Towards unified object detection and semantic segmentation. In *Proceedings of the European conference on Computer vision*. 2014.
- [26] J. Dong, W. Xia, Q. Chen, J. Feng, Z. Huang, and S. Yan. Subcategory-aware object classification. In *Conference on Computer Vision and Pattern Recognition*, 2013.
- [27] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012.
- [28] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Conference on Computer Vision and Pattern Recognition*, 2009.
- [29] P. Felzenszwalb and D. Huttenlocher. Distance transforms of sampled functions. In *Cornell Computing and Information Science*, 2004.
- [30] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Conference on Computer Vision and Pattern Recognition*, 2008.
- [31] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Cascade object detection with deformable part models. In *Conference on Computer Vision and Pattern Recognition*, 2010.
- [32] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Transactions on Pattern Analysis and Machine Intelligence*, 2010.
- [33] R. Feris, R. Bobbitt, L. Brown, and S. Pankanti. Attribute-based people search: Lessons learnt from a practical surveillance system. In *Proceedings*

- of *International Conference on Multimedia Retrieval*, 2014.
- [34] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *International Conference on Computer Vision*, 2013.
  - [35] V. Ferrari and A. Zisserman. Learning visual attributes. In *Advances in Neural Information Processing Systems*, 2008.
  - [36] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by image and video content: The qbic system. *Computer*, 1995.
  - [37] A. Gallagher and T. Chen. Clothing cosegmentation for recognizing people. In *Conference on Computer Vision and Pattern Recognition*, 2008.
  - [38] R. Girshick. Fast r-cnn. *arXiv preprint arXiv:1504.08083*, 2015.
  - [39] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Conference on Computer Vision and Pattern Recognition*, 2014.
  - [40] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *International Conference on Artificial Intelligence and Statistics*, 2011.
  - [41] B. Gong, F. Sha, and K. Grauman. Overcoming dataset bias: An unsupervised domain adaptation approach. In *Neural Information Processing Systems Workshop on Large Scale Visual Recognition and Retrieval*, 2012.
  - [42] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Conference on Computer Vision and Pattern Recognition*, 2012.
  - [43] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio. Maxout networks. *arXiv preprint arXiv:1302.4389*, 2013.
  - [44] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *International Conference on Computer Vision*, 2011.
  - [45] R. Gopalan, R. Li, and R. Chellappa. Unsupervised adaptation across domain shifts by generating intermediate data representations. *Transactions on Pattern Analysis and Machine Intelligence*, 2014.
  - [46] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative

- classification with sets of image features. In *International Conference on Computer Vision*, 2005.
- [47] J. He, J. Feng, X. Liu, T. Cheng, T.-H. Lin, H. Chung, and S.-F. Chang. Mobile product search with bag of hash bits and boundary reranking. In *Conference on Computer Vision and Pattern Recognition*, 2012.
  - [48] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.
  - [49] G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Improving neural networks by preventing coadaptation of feature detectors. 2012.
  - [50] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 1997.
  - [51] J. Hoffman, E. Tzeng, J. Donahue, Y. Jia, K. Saenko, and T. Darrell. One-shot adaptation of supervised deep convolutional models. *arXiv preprint arXiv:1312.6204*, 2013.
  - [52] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, 2007.
  - [53] J. Huang, S. Liu, J. Xing, T. Mei, and S. Yan. Circle & search: Attribute-aware shoe retrieval. *Transactions on Multimedia Computing, Communications, and Applications*, 2014.
  - [54] T. Huang, C. Dagli, and S. Rajaram. Active learning for interactive multimedia retrieval. *Proceedings of the ...*, 2008.
  - [55] V. Jagadeesh, R. Piramuthu, A. Bhardwaj, W. Di, and N. Sundaresan. Large scale visual recommendations from street fashion images. In *Proceedings of the international conference on Knowledge Discovery and Data mining*, 2014.
  - [56] E. Jaha and M. Nixon. Soft Biometrics for Subject Identification using Clothing Attributes. In *International Joint Conference on Biometrics*, 2014.
  - [57] H. Jégou, M. Douze, and C. Schmid. Product Quantization for Nearest Neighbor Search. *Transactions on Pattern Analysis and Machine Intelli-*

gence, 2011.

- [58] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *Computer Vision and Pattern Recognition*, 2010.
- [59] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, 2014.
- [60] Y. Kalantidis, L. Kennedy, and L. Li. Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos. In *Proceedings of the ACM Conference on International Conference on Multimedia Retrieval*, 2013.
- [61] H. Kang, M. Hebert, A. A. Efros, and T. Kanade. Connecting missing links: Object discovery from sparse observations using 5 million product images. In *Proceedings of the European Conference on Computer Vision*, 2012.
- [62] M. Kiapour, K. Yamaguchi, A. Berg, and T. Berg. Hipster Wars: Discovering Elements of Fashion Styles. In *Proceedings of the European Conference on Computer Vision*, 2014.
- [63] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Image search with relative attribute feedback. In *Conference on Computer Vision and Pattern Recognition*, 2012.
- [64] A. Krizhevsky and G. E. Hinton. Using very deep autoencoders for content-based image retrieval. In *European Symposium on Artificial Neural Networks*, 2011.
- [65] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.
- [66] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *Conference on Computer Vision and Pattern Recognition*, 2011.
- [67] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. Berg, and T. Berg.

- Baby Talk: Understanding and Generating Simple Image Descriptions. In *International Conference on Computer Vision*, 2011.
- [68] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *International Conference on Computer Vision*, 2009.
- [69] I. Kwak, A. Murillo, P. Belhumeur, D. Kriegman, and S. Belongie. From bikers to surfers: Visual recognition of urban tribes. In *British Machine Vision Conference*, 2013.
- [70] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Conference on Computer Vision and Pattern Recognition*, 2009.
- [71] R. Layne, T. Hospedales, and S. Gong. Person re-identification by attributes. In *British Machine Vision Conference*, 2012.
- [72] A. Li, L. Liu, K. Wang, S. Liu, and S. Yan. Clothing attributes assisted person re-identification. *Transactions on Circuits and Systems for Video Technology*, 2014.
- [73] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, L. Lin, and S. Yan. Deep human parsing with active template regression. *Transactions on Pattern Analysis and Machine Intelligence*, 2015.
- [74] M. Lin, Q. Chen, and S. Yan. Network in network. In *International Conference on Learning Representations*, 2014.
- [75] L. Liu, J. Xing, S. Liu, H. Xu, X. Zhou, and S. Yan. Wow! you are so beautiful today! *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2014.
- [76] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, and S. Yan. Fashion parsing with weak color-category labels. *IEEE Transactions on Multimedia*, 2014.
- [77] S. Liu, J. Feng, Z. Song, T. Zhang, H. Lu, C. Xu, and S. Yan. Hi, magic closet, tell me what to wear! In *Proceedings of the ACM international conference on Multimedia*, 2012.
- [78] S. Liu, X. Liang, L. Liu, X. Shen, J. Yang, C. Xu, L. Lin, X. Cao, and S. Yan. Matching-cnn meets knn: Quasi-parametric human parsing. *arXiv*

*preprint arXiv:1504.01220*, 2015.

- [79] S. Liu, L. Liu, and S. Yan. Fashion Analysis: Current Techniques and Future Directions. *Proceedings of the ACM International Conference on Multimedia*, 2014.
- [80] 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 *Conference on Computer Vision and Pattern Recognition*, 2012.
- [81] S. Liu, S. Yan, T. Zhang, C. Xu, J. Liu, and H. Lu. Weakly supervised graph propagation towards collective image parsing. *IEEE Transactions on Multimedia*, 2012.
- [82] B. Loni, L. Cheung, M. Riegler, A. Bozzon, L. Gottlieb, and M. Larson. Fashion 10000: an enriched social image dataset for fashion and clothing. In *Proceedings of the ACM Multimedia Systems Conference*, 2014.
- [83] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004.
- [84] S. Lu, T. Mei, J. Wang, J. Zhang, Z. Wang, D. D. Feng, J.-T. Sun, and S. Li. Browse-to-search. In *Proceedings of the ACM International Conference on Multimedia*, 2012.
- [85] P. Luo, X. Wang, and X. Tang. A Deep Sum-Product Architecture for Robust Facial Attributes Analysis. In *International Conference on Computer Vision*, 2013.
- [86] M. Manfredi, C. Grana, S. Calderara, and R. Cucchiara. A complete system for garment segmentation and color classification. In *Machine Vision and Applications*, 2013.
- [87] B. Neyshabur, R. Salakhutdinov, and N. Srebro. Path-sgd: Path-normalized optimization in deep neural networks. *arXiv preprint arXiv:1506.02617*, 2015.
- [88] H. V. Nguyen, H. T. Ho, V. M. Patel, and R. Chellappa. Joint hierarchical domain adaptation and feature learning. *Transactions on Pattern Analysis and Machine Intelligence*, 2013.
- [89] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern*

*Recognition*, 1996.

- [90] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks. In *Conference on Computer Vision and Pattern Recognition*, 2014.
- [91] D. Parikh and K. Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *Conference on Computer Vision and Pattern Recognition*, 2011.
- [92] D. Parikh and K. Grauman. Relative Attributes. In *International Conference on Computer Vision*, 2011.
- [93] D. Parikh and K. Grauman. Relative attributes. In *International Conference on Computer Vision*, 2011.
- [94] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Conference on Computer Vision and Pattern Recognition*, 2012.
- [95] F. Perronnin, J. Sanchez, and T. Mensink. Improving the Fisher Kernel for Large-Scale Image Classification. In *European Conference on Computer Vision*, 2010.
- [96] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. *Conference on Computer Vision and Pattern Recognition Workshops*, 2014.
- [97] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.
- [98] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele. What Helps Where And Why? Semantic Relatedness for Knowledge Transfer. In *Conference on Computer Vision and Pattern Recognition*, 2010.
- [99] R. G. J. S. Shaoqing Ren, Kaiming He. Faster R-CNN: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.
- [100] X. Shen, Z. Lin, J. Brandt, and Y. Wu. Mobile product image search by automatic query object extraction. In *Proceedings of the European Conference on Computer Vision*, 2012.

- [101] Z. Shi, T. Hospedales, and T. Xiang. Transferring a semantic representation for person re-identification and search. In *Conference on Computer Vision and Pattern Recognition*, 2015.
- [102] B. Siddiquie, R. S. Feris, and L. S. Davis. Image ranking and retrieval based on multi-attribute queries. In *Conference on Computer Vision and Pattern Recognition*, 2011.
- [103] Z. Song, M. Wang, X. Hua, , and S. Yan. Predicting occupation via human clothing and contexts. In *International Conference on Computer Vision*, 2011.
- [104] R. K. Srivastava, K. Greff, and J. Schmidhuber. Training very deep networks. In *Advances in Neural Information Processing Systems*, 2015.
- [105] Y. Sun, X. Wang, , and X. Tang. Deeply learned face representations are sparse, selective, and robust. In *Conference on Computer Vision and Pattern Recognition*, 2015.
- [106] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International journal of computer vision*, 2013.
- [107] K. van de Sande, J. Uijlings, T. Gevers, and A. W. M. Smeulders. Segmentation as Selective Search for Object Recognition. In *International Conference on Computer Vision*, 2011.
- [108] V. Vapnik and A. Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 2009.
- [109] D. A. Vaquero, R. S. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk. Attribute-based people search in surveillance environments. In *Workshop on Applications of Computer Vision*, 2009.
- [110] A. Vedaldi, S. Mahendran, S. Tsogkas, S. Maji, B. Girshick, J. Kannala, E. Rahtu, I. Kokkinos, M. Blaschko, D. Weiss, B. Taskar, K. Simonyan, N. Saphra, and S. Mohamed. Understanding Objects in Detail with Fine-grained Attributes. In *Conference on Computer Vision and Pattern Recognition*, 2014.
- [111] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. In *Technical Report*, 2011.



- [112] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li. Deep learning for content-based image retrieval: A comprehensive study. In *Proceedings of the ACM International Conference on Multimedia*, 2014.
- [113] J. Wang, J. Yang, K. Yu, F. Lv, and T. Huang. Locality-constrained linear coding for image classification. In *Conference on Computer Vision and Pattern Recognition*, 2010.
- [114] X. Wang and T. Zhang. Clothes search in consumer photos via color matching and attribute learning. In *Proceedings of the ACM International Conference on Multimedia*, 2011.
- [115] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. In *Proceedings of the European Conference on Computer Vision*, 2010.
- [116] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan. Cnn: Single-label to multi-label. *arXiv preprint arXiv:1406.5726*, 2014.
- [117] W. Xia, C. Domokos, , J. Xiong, L.-F. Cheong, and S. Yan. Segmentation over detection via optimal sparse reconstructions. *Trasactions on Circuits and Systems for Video Technology*, 2014.
- [118] W. Xia, C. Domokos, L. F. Cheong, and S. Yan. Background context augmented hypothesis graph for object segmentation. *Transactions on Circuits and Systems for Video Technology*, 2014.
- [119] W. Xia, C. Domokos, J. Dong, L.-F. Cheong, and S. Yan. Semantic segmentation without annotating segments. In *Proceedings of International Conference of Computer Vision*, 2013.
- [120] W. Xia, Z. Song, J. Feng, L. F. Cheong, and S. Yan. Segmentation over detection by coupled global and local sparse representations. In *Proceedings of European Conference of Computer Vision*, 2012.
- [121] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *Conference on Computer Vision and Pattern Recognition*, 2012.
- [122] J. Yang, K. Yu, and Y. Gong. Linear spatial pyramid matching using sparse coding for image classification. In *Conference on Computer Vision and Pattern Recognition*, 2009.

- [123] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Conference on Computer Vision and Pattern Recognition*, 2011.
- [124] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. PANDA: Pose Aligned Networks for Deep Attribute Modeling. In *Conference on Computer Vision and Pattern Recognition*, 2014.
- [125] J. Zhu, S. Liao, Z. Lei, D. Yi, and S. Li. Pedestrian attribute classification in surveillance: Database and evaluation. In *International Conference on Computer Vision Workshop on Large-Scale Video Search and Mining*, 2013.

# List of Publications

1. **Junshi Huang**, Rogerio Feris, Qiang Chen, Shuicheng Yan. Cross-domain Image Retrieval with a Dual Attribute-aware Ranking Network. In *International Conference on Computer Vision*, 2015
2. Qiang Chen\*, **Junshi Huang\***, Rogerio Feris, Lisa Brown, Jian Dong, Shuicheng Yan. Deep Domain Adaptation for Describing People Based on Fine-grained Clothing Attributes. In *Conference on Computer Vision and Pattern Recognition*, 2015.
3. **Junshi Huang**, Si Liu, Junliang Xing, Tao Mei, Shuicheng Yan. Circle & Search: Attribute-Aware Shoe Retrieval. In *ACM Transactions on Multimedia Computing, Communications and Applications*, 2014.
4. Junliang Xing, Zhiheng Niu, **Junshi Huang**, Weiming Hu, Shuicheng Yan. Towards Multi-view and Partially-Occluded Face Alignment. In *Conference on Computer Vision and Pattern Recognition*, 2014.
5. **Junshi Huang\***, Wei Xia\*, Shuicheng Yan. Deep Search with Attribute-aware Deep Network. In *Proceedings of the ACM International Conference on Multimedia*, 2014.
6. Yunchao Wei, Wei Xia, **Junshi Huang**, Bingbing Ni, Jian Dong, Yao Zhao, Shuicheng Yan. CNN: Single-label to Multi-label. *arXiv preprint arXiv:1406.5726*, 2014.
7. Si Liu, Jiashi Feng, Csaba Domokos, Hui Xu, **Junshi Huang**, Zhenzhen Hu, Shuicheng Yan. Fashion Parsing With Weak Color-Category Labels. In *IEEE Transactions on Multimedia*, 2013.

8. **Junshi Huang**, Hairong Liu, Jiale Shen, Shuicheng Yan. Towards efficient sparse coding for scalable image annotation. In *Proceedings of the ACM International Conference on Multimedia*, 2013.