On shoe attribute prediction and retrieval

Zhan, Huijing

2018

Zhan, H. (2018). On shoe attribute prediction and retrieval. Doctoral thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/73509>

<https://doi.org/10.32657/10356/73509>

On Shoe Attribute Prediction and Retrieval

Zhan Huijing

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirement for the degree of
Doctor of Philosophy

2017

To my families for their encouragements and love.

Acknowledgements

Immeasurable appreciation and deepest gratitude are extended to the persons below who have kindly provided extensive support over my past five-years study in Nanyang Technological University (NTU) and made it a wonderful as well as enriching experience.

First and foremost, I would like to take this opportunity to express my sincere thanks and gratitude to my supervisor, Professor Alex C. Kot, for his great support and generous mentorship during my PhD career. It is my great honour to be supervised by him and be his student. His continuing encouragement and patient support inspire me to keep on in front of challenges and devote all my efforts to overcome barriers, making my research career a wonderful and rewarding experience. Without him, I would never learn how to think deep and out of box. Without him, I would never have the courage to work on as a PhD student. I also remember the opportunities that he provided, such as the overseas exchange program to Tokyo Institute of Technology, competition participation and position of the project officer in the Rapid-Rich Object Search Lab (ROSE). His charming personal characteristics, professional attitude towards research, and passion for life have greatly promoted me in both my personal and academic life. Without his extensive and professional knowledge, persistent mentoring and support, this work would never have been possible.

I am also grateful to many professionals for their kind helps and guidance. Special thanks should be given to Dr. Shi Boxin for his patience in teaching me the paper writing skills, formula deduction skills and presentation tips. With his kind help, I learn how to conduct independent

research and how to organize a paper nicely. Moreover, I would express my sincere thanks for his contribution to the fourth and fifth chapter of the thesis. In addition, I would like to extend my thanks to my TAC panel members, Professor Yuan Junsong, Prof Xudong (now at the University of Sydney), and Prof Lin Weisi for their kindness, insightful suggestions and technical discussions. I would also like to thank Dr. Li Sheng not only for discussions and collaborations but also for helping me to improve the scientific writing and contributing to the third chapter of the thesis. I would also like to acknowledge Dr. Wangyan and Dr. Wang Anran, for their tremendous help and encouragement. I am greatly appreciated for their support and encouraging words during my frequent moments of stress. I am thankful to the Deputy Director of ROSE lab, Dr. Dennis Sng, the Principal Laboratory Executive of ROSE lab, Mr. Joseph, and the Assistant Center Manager, Mrs. Wang Qian for offering me the student helper position and countless help in trivial lab issues. Many thanks are also owing to all my colleagues in ROSE Lab and Workstation Resource Laboratory. I would not be able to make rapid progress without the help from my senior Dr. Yang Huijuan, Dr. Lai Jian, Dr. Fan Jiayuan, Dr. Liu Siyuan, Dr. Miao Zhenwei, Dr. Chen Changsheng, Dr. Baiying Lei and Dr. Mu Hao. The labs would not be so harmonious without the work of Mr. Steward Chu, Mr. Joseph Lim, and Mrs. Wang Qian. I also enjoy the time with all the group members and friends, Lu Ze, Li Haoliang, Wan Renjie, Devadeep, Tra, Terry, Abrar, Hu Junlin, Weng Renliang, Zhang Tianyi, Zhang Yabin, Zhang Xinfeng, Zhang Dajiang, Meng Jingjing, Amir, Wang Zhe, Yang Jiong, Ye Wei, Li Qian, Li Bingbing, Wang Xiaoning, Tang Chao.

I am deeply in debt to my parents (Wan Honghua and Zhan Yulong) for instilling me the perseverance and hard-working from childhood which benefit me no matter what I choose to do. I am grateful to my sister, Mrs. Wan Wei for her inspiration and encouragements. My dream towards a PhD can never be possible without the strong supports from my family.

Contents

Acknowledgements	1
Abstract	6
List of Figures	8
List of Tables	11
List of Abbreviations	13
1 Introduction	1
1.1 Background and Motivation	1
1.2 Objectives and Major Contributions of the Thesis	7
1.3 Organization of the Thesis	10
2 Literature Review	11
2.1 Describing Objects with Attributes	11
2.1.1 What is attribute?	11
2.1.2 Attribute-based Visual Representation	17
2.2 Instance-level Image Retrieval	19
2.2.1 Feature Representation for Instance Retrieval	20
2.2.2 Metric Learning for Instance Retrieval	26
2.2.3 Image Retrieval for Fashion Items	29
2.3 Multi-Task Learning	31
2.4 Object Detection	32
2.5 Summary and discussions	35

3	Towards Multi-View Shoe Image Indexing and Retrieval	37
3.1	Introduction	38
3.2	Multi-View Shoe Dataset	41
3.2.1	Multi-view Shoe Images Collection	41
3.2.2	Annotation of shoe part locations and attributes	42
3.3	The Proposed System	43
3.3.1	System Overview	43
3.3.2	View Identification	44
3.3.3	Relevant View Selection	45
3.3.4	View-specific Part Localization	45
3.3.5	Feature Extraction and Classification	47
3.3.6	Index distance measure	47
3.4	Experimental Results and Discussions	48
3.4.1	Performance of Viewpoint Identification	49
3.4.2	Performance of Attribute Estimator	49
3.4.3	Performance of Shoe Retrieval	53
3.4.4	Discussions	56
3.5	Summary	56
4	Cross-Domain Shoe Retrieval with a Semantic Hierarchy of Attribute Classification Network	58
4.1	Introduction	59
4.2	Our Approach	64
4.2.1	SHOE-CNN: Semantic Hierarchy of attribute based Convolutional Neural Network	64
4.2.2	Three-Level Feature Representations for Shoes	69
4.2.3	Deep Features Activated from three-Level SHOE-CNN	71
4.2.4	Similarity Metric Learning	72
4.3	Dataset Construction	74
4.4	Experimental Results	76
4.4.1	Data Preparation and Implementation Details	76
4.4.2	Evaluations of the Proposed System	79
4.5	Summary	88

5	DeepShoe: A Multi-Task View-Invariant CNN for Model-to-Shop Shoe Retrieval	89
5.1	Introduction	90
5.2	Model-to-Shop Shoe Retrieval	93
5.2.1	The Overall Framework	93
5.2.2	Model-to-Shop Shoe Retrieval Using Triplet Network	95
5.2.3	Multi-Task View-Invariant CNN	97
5.2.4	Attribute-based Hard Triplet Weighting Strategy	102
5.3	Dataset Construction	106
5.4	Implementation Details	107
5.5	Experiment Settings	109
5.5.1	Training and Testing Data	109
5.5.2	Comparison Methods	110
5.5.3	Evaluation Protocol	110
5.5.4	Results	111
5.6	Summary	119
6	Conclusions and Future Work	121
6.1	Conclusions	121
6.2	Future Work	123
	Author's Publications	126
	References	127

Abstract

With the rapid proliferation of Internet, it has become a great challenge to annotate a large number of objects manually, especially for the fashion domain where a massive collection of new products come up everyday. Moreover, the huge profits brought by the online fashion market have motivated the multimedia search using the vision-based techniques. Therefore, to save human labor, it is essential to develop an automatic tagging system (i.e., an attribute prediction system) for those fashion products in a variety of appearances. With the annotated attributes, it enables the online shoppers to perform the retrieval of their desired shoes. However, it fails to work on the shoe images from the daily life because their large visual difference from the images of online stores leads to unsatisfactory attribute prediction performance. In this thesis, we study the shoe attribute prediction and retrieval for the online images as well as the daily life photos. More specifically, we address the problem of in-store shoe retrieval as well as the more challenging issue of cross-scenario shoe retrieval guided by the semantic attributes of shoes. The works in this thesis can be summarized as below.

An in-store shoe image indexing and retrieval system which allows the query images in the form of multi-view shoe images is firstly proposed here. Given a set of multi-view shoe images, we first identify the viewpoint of each shoe image and a set of relevant view images are selected to estimate the value of each shoe attribute. To effectively predict the attributes which are part-aware, we incorporate the prior knowledge of the shoe structure under a certain viewpoint to learn a novel view-specific part localization model. Similarly, each shoe in the database

is also indexed by a list of part-aware shoe attributes. Experimental results demonstrate the effectiveness of the proposed system on a newly-built structured multi-view online shoe dataset.

Not limited to the attribute prediction and retrieval for the online store shoe images, we also relax the constraint to perform the same task for the daily life photos with cluttered backgrounds, different scales, varied viewpoints, etc. A novel cross-domain shoe retrieval system is presented which aims to find the exactly same shoes given the query daily life shoe photos. More specifically, we propose the Semantic Hierarchy Of attributE Convolutional Neural Network (SHOE-CNN) with a newly designed loss function which systematically merges semantic attributes of closer visual appearances to avoid shoe images with the obvious visual differences being matched wrongly. Moreover, a coarse-to-fine three-level feature representation is developed to effectively match the shoe images across different domains. The experimental results demonstrate the advantages of each component of our proposed system and a significant improvement over other baseline methods.

To further enable the retrieval of online store images with different viewpoints and address the failure cases with large viewpoint variation while at the same time improving the capability of differentiating the fine-grained details, we propose the feature embedding for shoes via a multi-task view-invariant convolutional neural network (MTV-CNN), the feature activations of which reflect the inherent similarity between any two shoe images. Specifically, we propose 1) the weighted triplet loss to reduce the feature distance between the same shoes in different scenarios; 2) a novel view-invariant loss to reduce ambiguous feature representation from different views; 3) a novel definition of shoe style based on combinations of part-aware semantic shoe attributes and the corresponding style identification loss are presented; and 4) the attribute-based hard negative and anchor images mining process to distinguish fine-grained differences. The experiments conducted on our newly collected dataset indicate that we are capable of not only returning the exactly same shoes or similar shoes but also different viewpoint ones.

List of Figures

1.1	The current and predicted sizes of the online fashion markets of USA, Europe and China in 2016-2021. Images from [1].	2
1.2	The illustration for the in-store shoe retrieval system (a) and cross-domain shoe retrieval system (b). In (a), the leftmost dashed black box indicates the ground truth attributes for the query image from the online shops. The green stars and the red crosses represent the correctly and wrongly matched attributes, respectively. The query image in (b) is from the daily life scenario and the bounding box with the green check mark indicates the exactly matched one as the query. .	3
1.3	Example multi-view images for the landmarks and shoes.	7
2.1	The typical pipeline for training the attribute classifier. Images from [41]. . . .	12
2.2	The concept for the semantic attribute (a) and discriminative attribute (b). . . .	13
2.3	The example images with the binary attribute annotations for sport shoes, face, and chair, respectively.	14
2.4	The binary attribute encodes the presence or absence of the “casual style” attribute with a bit. It is easy to define that (a) is casual while (c) is formal, not casual. However, in some cases it is difficult to draw a clear line about whether this shoe is casual or not, as shown in (b).	15
2.5	Several examples for the local fine-grained semantic attributes for the dress, bird, and shoes. The localizations for the attributes are highlighted in the green and yellow bounding box for the associated attributes in the same font color. . .	16
2.6	(a) The concatenation of attribute prediction scores and (b) attribute prediction guided visual feature learning.	17
2.7	The block diagram of a typical instance-level image retrieval system. Given a query image depicting a particular object (we take the shoes as a concrete example), instance-level image retrieval aims at finding the images containing exactly the same object in the reference database.	20
2.8	Example network architectures for AlexNet [75] and VGGNet [130]. Images from [161].	22
2.9	The architecture of the AlexNet with Images from [76].	23

2.10	The pipeline of object detection framework utilizing the region-proposal based object detection utilizing the ranking-based methods and classifier-based methods.	34
3.1	Examples of common viewpoints of a shoe and the manually annotated parts for shoe images in our dataset (best viewed in pdf file and color images).	41
3.2	Illustrations of the appearance for each shoe attribute. The first row of images correspond to the presence of each attribute, while the second row indicates the absence of the attributes.	42
3.3	The proposed retrieval system with multi-view query shoe images.	44
3.4	Part localization results of the proposed view-specific part localization approach. In the result, the bounding box in different colors represent the detected shoe parts. (a) single view shoe images. (b) five viewpoint images for heeled shoes with different heights.	51
3.5	Performance of shoe retrieval (in terms of NDCG scores) using our proposed system and the state-of-the-art system [64].	54
3.6	Comparison of retrieval results using our proposed system and the state-of-the-art system [64].	55
3.7	Examples of shoe image tagging.	56
4.1	Example street domain and their counterpart online domain shoe images. The street domain photos may have scale variation, cluttered background, human feet, different poses and viewpoints of shoes, as well as illumination variation. In contrast, pictures from online shops are captured in clean background, ideal illumination conditions as well as fixed viewpoints.	60
4.2	The hierarchical structure of semantic shoe attributes (in black box) and subordinate attributes (SA, in red box) for toes and heels. By grouping “Square Toe”, “Round Toe”, and “Pointy Toe” as “Closed Toe” (SA1) and “Peep Toe” as “Open Toe” (SA2), the semantic attribute prediction will hardly confuse shoes with visually dissimilar toes. The appearances of the attributes are illustrated in (a). The same applies to heels, as shown in (b).	61
4.3	The overview of our proposed cross-domain exact shoe retrieval system (best viewed in color).	63
4.4	The architecture of our SHOE-CNN network. It consists of five convolution layers and one fully-connected layers shared for all the attributes. However, FC1 branches out T attribute-specific layers (T is the number of semantic shoe attribute types) thus each attribute owns independent FC2 and FC3 layers.	65
4.5	The 1st column pictures indicate the query real-world shoe photos and the right-most three columns illustrate the top-3 region proposals selected by our proposed method. The middle part shows the initial coarse region proposals generated by the EdgeBox algorithm [179], and the top-3 scored candidates are highlighted in green.	69

4.6	The Image collage from our complete dataset: Left picture demonstrates the images from the street domain and right picture shows the photos from the online domain. Example corresponding images of exact the same pair of shoes from both domains are illustrated in close-up views. Each pair of shoes are represented using three images from the street domain (highlighted in red boxes) and its counterpart online shoe images (highlighted in green boxes). Please zoom-in the electronic version for better details.	77
4.7	Example retrieval results with top-8 returned shoe candidates. The correctly retrieved results with exactly the same shoe to the query are highlighted in green.	84
4.8	SHOE-CNN feature maps. Row 1 and 3: two examples of the daily shoe photos (wedge and high-heel shoes). Row 2 and 4: corresponding activated feature maps from the pool5 layer for the whole query image (Level 1), the corresponding top-scored region proposals (Level 2) referring to each single shoe of a pair, and the local fine-grained patches (Level 3).	86
4.9	(a) Detection performance with different number of selected region proposals. (b) Mean average attribute prediction accuracy with respect to different values of λ . (c) Top-20 retrieval accuracy when we vary the parameter γ	86
5.1	Illustrations of main challenges lie in the model-to-shop shoe retrieval. The top two rows (Challenge I) show pairs of exact-match model and online store shoe images with scale, viewpoint, illumination and occlusion changes. The middle row (Challenge II) demonstrates the different shoes with very small fine-grained differences. In the last row (Challenge III), the leftmost column of (c) is the query daily shoe photo, we can retrieve the shoes in the green dashed box with similar viewpoints as that of the query image. However, the system fails at finding the shoe images in the red dashed box in less similar viewpoints.	91
5.2	The framework of our model-to-shop shoe retrieval. Here a, p, n indicate the 512-d feature vectors for the anchor, positive and negative image, respectively. The feature representations for the different viewpoint images of the anchor image, named as positive bag, are denoted as p_1, p_2, p_3	94
5.3	The basic architecture of the traditional triplet network with three images (x_i^a, x_i^p, x_j^n) as the input.	97
5.4	The illustration for the view-invariant loss. In the right figure, the positive examples of the same shoes (green box) are pulled closer. By minimizing the view-invariant loss as an auxiliary task, the anchor (blue box) and positive bag (green box) are clustered while the negative (red box) are pushed away.	102
5.5	Hierarchical grouping of the tree-structured semantic shoe attributes for “Toe Shape”.	104
5.6	Examples of cropped model and shop scenarios images in our MVShoe dataset.	108
5.7	Top@K retrieval accuracy with K varying from 1 to 80.	115
5.8	Example retrieval results shown with the green bounding box indicates the exact-match to the query item.	115

List of Tables

3.1	Part-aware attributes and the corresponding relevant views.	42
3.2	The visible parts of each viewpoint.	43
3.3	Shoe part localization accuracy (in terms of PCP metrics) using our proposed method (shown in bold number) and the work in [64] (shown in non-bold number).	49
3.4	Attribute prediction accuracy of our proposed method (shown in bold number) and the work in [64] (shown in non-bold number) for different views.	50
3.5	The performance of attribute estimation on the “DailyShoes” dataset using the proposed method and the state-of-the-art [64].	51
3.6	The performance of attribute estimation for a set of M views of each shoe using our proposed method (shown in bold number) and the comparison with [64] when $M = 3$ (shown in non-bold number).	52
3.7	The performance of attribute estimation with the deep features compared to that with HOG feature.	52
3.8	The subjective evaluation of our proposed retrieval system with a set of M views for each shoe (shown in bold number) and the comparison with [64] (shown in non-bold number).	56
4.1	Example attribute values in our dataset. The numbers in parenthesis indicate the total number of values for the corresponding attribute types.	76
4.2	Attribute prediction of the proposed SHOE-CNN compared with other approaches and different settings of subordinate attributes.	81
4.3	Comparison of the multi-class attribute prediction accuracy on clothing attribute dataset [26].	82
4.4	Shoe detection performance of real-world shoe photos comparing with two baselines: CNN detection and DPM detection.	82
4.5	Top- K Retrieval Accuracy of our system and baseline results.	83
5.1	Semantic Attribute Groups in MVShoe dataset.	108
5.2	Top-20 retrieval accuracy on the MVShoe dataset.	112

5.3	NDCG score on the MVShoe dataset.	113
5.4	Comparisons of different CNN architectures on the retrieval performance in terms of Top@20 Accuracy when utilizing the ratio loss.	117
5.5	Comparisons of different CNN architectures on the retrieval performance in terms of Top@20 Accuracy when utilizing the margin loss.	117
5.6	Contributions of the auxiliary tasks and each stage to the proposed MTV-CNN for the ratio loss.	118
5.7	Contributions of the auxiliary tasks and each stage to the proposed MTV-CNN for the margin loss.	119

List of Abbreviations

CBIR	Content-based image retrieval
DoG	Difference of gaussians
MSER	Maximally stable extremal region
FV	Fisher vector
DPM	Deformable part model
HOG	Histogram of gradient
SGD	Stochastic
SVM	Support vector machine
NDCG	Normalized discounted cumulative gain
CNN	Convolutional neural network
SHOE-CNN	Semantic hierarchy of attribute convolutional neural network
LBP	Local binary pattern
Conv	Convolutional
FC	Fully-connected
YOLO	You Only Look Once
SSD	Single Shot MultiBox Detector
DSSD	Deconvolutional Single Shot Detector
Positive Semi-Definite	PSD
OASIS	Online algorithm for scalable image similarity learning
VLAD	Vector of locally aggregated descriptors
POOFs	Part based one-vs-one features
CRF	Conditional random field
IoU	Intersection-over-union
SPP	Spatial pyramid pooling
MAP	Mean average precision
SA	Subordinate attribute
PCA	Principle components analysis
DSIFT	Dense scale-invariant feature transform
MTV-CNN	Multi-task view-invariant convolutional neural network
MVShoe	Multi-view shoe dataset with semantic attributes

SURF	Speed-up robust features
LMNN	Large margin nearest neighbor
ITML	Information-theoretic metric learning
CCL	Coupled cluster loss
MTL	Multi-task learning
ROIs	Region of interests
TN	Triplet network
AP	Attribute prediction
SS	Style identification
VI	Viewpoint invariance
R-CNN	Regions with convolutional neural network features
SIFT	Scale-invariant feature transform
SGD	Stochastic gradient descend
GANs	Generative adversarial networks

Chapter 1

Introduction

1.1 Background and Motivation

With the widespread availability of the Internet, online shopping industry has been expanding rapidly in these years and dramatically changed the consumers' purchasing behaviour. Compared to the brick and mortar store shopping, the novel online shopping experience not only facilitates the customers to gather more detailed information about their desired products (e.g., size, texture, color, consumer ratings and feedbacks), but also provides more choices of goods at better prices, which are quite attractive for consumers. It is reported that in 2016 about 1.61 billion people worldwide purchase goods online and the global e-commerce sales is amounted to about 1.9 trillion U.S.dollars [3]. Among various categories of products, the online fashion market makes up a large portion of the overall e-commerce sales and it is estimated to reach about 530 billion [1] in the next four years, as shown in Fig. 1.1. However, behind the huge profits, online sellers need to take great efforts to maintain as well as update their products. For instance, they need to update the fashion images of the displayed products frequently as well as the corresponding descriptions. Nevertheless, manually generating these descriptions is costly, which motivates us to design an automatic product description annotation system (i.e., the attribute prediction system). Note that in the rest of the thesis, "shoe indexing", "automatic tagging" and "semantic attribute prediction" are utilized interchangeably.

Fig. 1.1: The current and predicted sizes of the online fashion markets of USA, Europe and China in 2016-2021. Images from [1].

Moreover, it often occurs in our online shopping experience when we are browsing for the desired product, we would like to find similar items as that one as well. We call it as *in-store retrieval* (see Fig. 1.2 (a)). Given a query fashion photo from the online store, with the product in the center of the image, the goal is to retrieve similar images from the reference gallery. Here one may ask that is it possible to search for the exact or similar product from online stores if the query image is from the daily life? Assuming that we are shopping in the real stores or walking on the street, and see a beautiful fashion product, we might take a picture of it and search through the website to find exactly the same or similar item with better prices. This problem is called *cross-domain retrieval*: the query and reference images are from different domains (see Fig. 1.2 (b)). The cross-domain retrieval system is considered as a more challenging case than the in-store retrieval system in the sense that the query image comes from the daily life with cluttered backgrounds, large variations of scale, viewpoints, etc.

For these two problems, the vision-based multimedia retrieval system is capable of finding the desired items in both the in-store and cross-domain scenarios. The approaches for image retrieval can be generalized into two categories: text-based and content-based image retrieval. The former returns the image list according to the annotated tags; however, the assumption

Fig. 1.2: The illustration for the in-store shoe retrieval system (a) and cross-domain shoe retrieval system (b). In (a), the leftmost dashed black box indicates the ground truth attributes for the query image from the online shops. The green stars and the red crosses represent the correctly and wrongly matched attributes, respectively. The query image in (b) is from the daily life scenario and the bounding box with the green check mark indicates the exactly matched one as the query.

is that the image should be labelled with descriptions, which might not be always available. Besides that, it poses difficulties for the consumers to describe their desired items with several key words due to the limited descriptive capability of words. The latter deals with the visual content of the products (i.e., the image), which can depict the details of the products without any distorted interpretation. The general content-based image retrieval engine (e.g., Google Image) is able to return related images according to the query image. Nevertheless, it usually fails to present satisfying results for the fine-grained fashion products (e.g., clothes, bags or shoes), and a subtle difference can even change the appearances of the objects. The examples include short sleeve or long sleeve for clothes, and high heel or flat heel for shoes.

Moreover, over the past decade, compared to the content-based image retrieval system at the category level, less attention is drawn to that at the instance level (i.e., to retrieve the exactly same item), especially for the fashion items. Therefore, it is essential to develop a fine-grained product image based retrieval system.

Shoe is a representative category in the fashion area. According to IBISWorlds report [2], the shoe store industry is expected to flourish in the coming years and surpass brick-and-mortar retailers. Moreover, compared with 33.4 percent in 2010, consumers earning over \$120,000 per year accounted for 42.6 percent for the expenses on shoes and online shoe sales was increased at an annualized rate of 15.2 percent to \$12bn over the past six years to 2016 [2]. Thus, this huge revenue in the online shoe shopping motivates our research in this specific fashion type. In this thesis, we focus on the semantic shoe attribute prediction and image-based shoe retrieval including the in-store shoe retrieval and the cross-domain shoe retrieval.

Predicting semantic shoe attributes and shoe retrieval is a novel instance of the fine-grained object attribute prediction and retrieval that are not well addressed before. In recent years, there is an increasing trend to develop efficient semantic attribute prediction and instance retrieval methods, which are summarized as below:

1. Semantic attribute prediction: aiming at describing the objects by a set of visual attributes (e.g., color, shape). The semantic attribute is supposed to cross the barrier of different categories and acts as a connection to bridge the gap between the human perception and computer understanding. It has attracted extensive attention from researchers in the multimedia community attributing to its successful application in the generic object recognition, zero-shot recognition, etc. The task of fine-grained classification greatly benefits from the introduction of the fine-grained attribute coupled with the local part (i.e., color at the wing, shape of the head, etc), which is supposed to describe the object part in detail and captures the subtle visual differences between different subcategories of a basic category (e.g., bird, flower, aircraft). Emerging

works begin to incorporate the semantic attributes into the task of fine-grained retrieval (i.e., clothing retrieval [26, 27, 63], fine-grained sketch retrieval [83, 160]) and achieve satisfactory performances. Traditional attribute prediction pipeline [46, 79, 82] generally extracts the hand engineered features of the images, based on which the off-the-shelf classifier is learnt for each attribute type. To improve the prediction accuracy of part-aware semantic attribute, existing approaches firstly localize the part, with which the particular attribute is correlated, and then extract the features pooled at the part region, such as facial attribute prediction [16, 79] and attribute prediction for birds [16, 171]. Recently, deep Convolutional Neural Network (CNN) has been introduced to address the problem of the semantic attribute prediction and demonstrated significant improvements [39, 94, 96, 172] over traditional pipeline. Rather than treating the prediction of various attributes as an independent and isolated process, there are some works investigating the correlation relationship between different types of attributes (e.g., mutual dependencies of the clothing attributes [26], as well as the hidden correlation between the facial attributes in a general context of subjects [134]).

2. Instance retrieval: aiming at returning the images of a specific object given a query example image of that particular object. The state-of-the-art instance retrieval methods describe the content of the object in terms of the following three directions: 1) describing the image by low-level hand-crafted local features and large visual codebooks employing Bag-of-Words (BoW) [132] algorithm; 2) developing a more compact descriptor by encoding aggregated low-level features with Fisher Vector [107], VLAD [69], etc; 3) deep features activated from the pre-trained CNN using large-scale dataset (e.g., ImageNet [34]) or extracted from the fine-tuned CNN using the task-oriented dataset. Besides an effective feature representation for the object, the determination of an appropriate distance metric is also critical in designing an instance retrieval system. As opposed to directly employing the pre-defined distance metric to evaluate the image similarity, the metric learning methods make the full use of the available side information

to incorporate the high-level knowledge. Here, the side information can be the pairwise relations between objects, for instance, whether two objects are similar or not in terms of category labels. Recent works [28, 57] demonstrate that it is possible to jointly learn a robust feature representation and an effective metric learning approach under the framework of CNN. The joint feature representation and metric learning technique are especially useful in addressing the problem of cross-domain retrieval such as person re-identification [28] when the query image and reference image are captured across multiple cameras with distinct viewpoints, and the consumer-to-shop product retrieval [57] when the query image possesses cluttered background while the reference images are with white background. The core idea is to learn a semantic and discriminative embedding space, in which the images of the same object are pulled closer and simultaneously the images of different objects are pushed away. In this way, the system is more likely to return the candidate retrieval list that is consistent with the human perception of similarity.

The above mentioned techniques are promising in some aspects, however, there are still many issues needed to be handled. First of all, the online store images of a particular shoe item are usually shown in multiple viewpoints, from which the appearances of the shoe might be quite different as well as the feature representation of images under those viewpoints. If we follow the conventional pipeline to develop a single SVM classifier for each attribute, it might lead to unreliable attribute prediction results. In addition, the CNN-based multi-class attribute prediction method treats different semantic attributes equally without exploring the inherent properties of the multi-class semantic attributes of shoes.

Due to the unique challenge of shoes from the large viewpoint variation (see Fig. 1.3), most of the instance retrieval methods [8, 71] achieving excellent performances on one-sided or near-planar objects [135] (e.g., buildings, logos and scenes) might lose its power on the shoe retrieval. The cross-domain shoe retrieval is more challenging than the in-store shoe retrieval

Fig. 1.3: Example multi-view images for the landmarks and shoes.

due to the large visual discrepancies between the query image in the daily life scenario and the reference image from the online store. The conventional instance retrieval methods don't take the domain difference into consideration. Moreover, the in-store and cross-domain shoe retrieval are fine-grained object retrieval, where different shoes belonging to the same style might differ from each other in minor fine-grained details. Deep feature descriptors (e.g., conv5 or fc6) with no semantic information supervision are not expressive to capture such differences. More importantly, the performance of joint feature representation and metric learning approach (e.g., triplet network) highly depends on how we select hard examples for efficient parameter learning. The current hard example mining algorithms are mainly focused on the hard negative example with less focus on the hard positive example, which is of equal importance.

1.2 Objectives and Major Contributions of the Thesis

The research carried out in this thesis focuses on shoe attribute prediction and retrieval. The objective lies in three aspects: (1) predicting the shoes attributes for the online store images and daily photos; (2) in-store shoe retrieval given multiple clean query shoe images facing different views; 3) cross-domain shoe retrieval, given a daily life photo depicting the shoe item.

The major contributions of this thesis can be summarized as follows:

1. Semantic shoe attribute prediction and in-store shoe retrieval based on multi-view shoe images: a novel attribute-based shoe image indexing and in-store retrieval system based on multi-view shoe images is developed. To the best of our knowledge, we are the first to take advantage of the complementary information that the multi-view images of shoes provide, and utilize this multi-view information to support the semantic attribute prediction and in-store shoe retrieval. For a particular attribute, its value is determined by the attribute prediction results over several relevant views, from which the appearance of the attribute is visible. Based on the prior knowledge of shoes appearance under a certain view, we propose a simple but effective view-specific part localization technique to localize the part that is associated with the particular attribute. Finally, each shoe item is represented by several bits indicating the presence or absence of the attributes, which largely facilitates the process of the in-store shoe retrieval. We conduct the experiments on two newly built shoe databases, one consists of 35670 multi-view shoe images collected from online merchants, and the other contains 930 daily shoe images collected from social websites. The experimental results on the shoe attribute estimation improve 4% in accuracy when compared with the state-of-the-art system. We also show the advantage of using multi-view shoe images for indexing and retrieval. It can be seen that even though the approach is designed for the in-store shoe retrieval, it is also applied to daily shoe photos with relatively simple background.

2. Cross-domain shoe retrieval with semantic hierarchy of attribute classification network: a cross-domain shoe retrieval framework is designed, which aims at finding the exact same shoe image in online shops given a daily life query image. We investigate the semantic hierarchy of attributes and propose the subordinate attribute to ensure that the dominant visual dissimilarity on shoe images is more reliably retained. Moreover, we propose a novel Semantic Hierarchy Of attribute Convolutional Neural Network (SHOE-CNN) with the newly designed

loss function to benefit the effective retrieval feature learning. To the best of our knowledge, we are the first to explore the intrinsic hierarchical properties of semantic attributes of shoes for cross-domain instance retrieval. Moreover, we develop a three-level feature representation to effectively extract the multi-scale information from shoe images and combine them with the multi-layer features of SHOE-CNN to conquer the unique challenges of cross-domain shoe retrieval. We also contribute a large-scale shoe dataset with images from both the daily life and the online store labelled with fine-grained attributes. The experimental results performed on the newly built dataset demonstrate that the proposed method outperforms the state-of-the-art instance retrieval methods and baseline CNN-based descriptors by a large margin.

3. A Multi-Task View-Invariant CNN for Model-to-Shop Shoe Retrieval: We extend the previous work by dealing with a more challenging case when the reference images are in different viewpoints, which is more close to the real-world searching scenario. Also, we aim to improve the capability of differentiating the fine-grained details of shoes as well as dealing with the failure case with large viewpoint variation. To this end, we develop a multi-task view-invariant CNN based on the triplet network to learn an effective feature representation for shoes, which reduces the domain gap by minimizing the distances between daily images and online store clean images. Moreover, to incorporate the semantic information into the learnt feature representation, we design an attribute-based hard example weighting and mining strategy to differentiate the subtle differences in appearances. To return more images of the exactly matched shoe item and mitigate the mismatching caused by the viewpoint difference, we treat the different view images of the same shoes as a positive bag and their mutual distances are minimized to pull them closer. We expand the previous dataset to include the multi-view images for each shoe item and collect more categories such as sport shoes, boots, etc. The experiments performed on our newly built dataset show our excellence when compared with other methods. We also examine the importance of different triplet loss functions in the performance of the cross-scenario shoe retrieval system.

1.3 Organization of the Thesis

This thesis is organized as follows.

- In Chapter 2, we present the detailed literature reviews for the the relevant fields including attribute-based representation, instance-level retrieval and object detection. For each topic, we first introduce the traditional approach utilized followed by the CNN-based methods. Moreover, we discuss some existing works about the image retrieval on the fashion products.
- In Chapter 3, we develop a novel attribute based shoe indexing and retrieval system based on multi-view shoe images, of which a set of relevant views are selected for predicting each part-aware semantic shoe attribute and the index of the shoe item is computed for in-store shoe retrieval.
- In Chapter 4, we demonstrate a cross-domain shoe retrieval framework and introduce the Semantic Hierarchy Of attributE Convolutional Neural Network (SHOE-CNN) with a newly designed loss function investigating the hierarchical properties of shoe attributes to improve the retrieval accuracy by extracting the multi-level deep feature representation for shoes.
- In Chapter 5, we extend the work in Chapter 4 to a more challenging case when the reference gallery contains the online clean images with multiple views. A multi-task view-invariant CNN is proposed to obtain a more effective and robust feature representation for the cross-domain shoe retrieval, which is supposed to be domain-invariant, semantic and discriminative as well as view-invariant.
- In Chapter 6, conclusions are drawn and future research directions are suggested.

Chapter 2

Literature Review

In this chapter, we provide a comprehensive literature review about the related topics, including the attribute-based representation, instance-level image retrieval, multi-task learning and the techniques for the object detection. We also discuss the difference between the methods proposed in this thesis and the related works.

2.1 Describing Objects with Attributes

In this section, we first describe the definition of the attribute followed by different types of attributes including the binary attribute, relative attribute, multi-class attribute as well as the fine-grained attribute. Then the related works based on the attribute-based representation are presented and analyzed.

2.1.1 What is attribute?

For the human visual system, the perception of an object is built by extracting its high-level semantics and characterizing the object with the descriptions such as “this wheel is round” or “this doll is furry”. Here “round” and “furry” are the concepts that human utilized to depict the visual properties of the object. However, these concepts easily interpreted by human cannot be

Fig. 2.1: The typical pipeline for training the attribute classifier. Images from [41].

understood by computer, which understands the object by analyzing the image content through pixels. Thus, there exists a semantic gap between human understanding and machine analysis. To address this discrepancy, Farhadi et al. [41] firstly proposed the term of “semantic attribute” to describe the properties of the object (e.g., “silky” for the materials, “pointy” for the shape, etc) and learnt the attribute classifiers with discriminatively selected base features (e.g., color, visual words, texture) to predict the presence or the absence of the particular attribute. Finally, the predicted attributes are stacked as more semantic and compact features to recognize the object category, as shown in Fig. 2.1. As opposed to previous works on naming the object category, describing the object by attributes enables us to acquire more information about the appearance of the objects. Moreover, we can utilize the attribute to not only describe the familiar objects but also generalize the concept to the unseen object.

The “semantic attribute” (see Fig. 2.2 (a)) enables us to describe the objects by textual words, however, they are not always sufficient to distinguish between different categories. To address this issue, there comes another type of attributes which can’t be explicitly defined by words, namely “discriminative attribute” (see Fig. 2.2 (b)). They correspond to the random splits in the visual space [41]. For example, cats have certain visual traits but dogs don’t have. In this thesis,

Fig. 2.2: The concept for the semantic attribute (a) and discriminative attribute (b).

we more focus on the “semantic attribute” that are human-defined and in the rest of the thesis, “attribute” is referred to as the “semantic attribute” for abbreviation.

2.1.1.1 Binary Attribute

Due to the aforementioned advantages of attribute-based representation and its promising performance in [41], the concept of attribute has drawn extensive attention in the multimedia community and has been applied to a variety of computer vision tasks such as zero-shot image classification [5, 6, 82, 146], face verification [15, 78, 79, 134], action recognition [48, 89, 116, 155, 158], image search [42, 63, 78, 127], etc. Binary attribute-based representation utilizes “yes” or “no” to indicate whether the given image demonstrates the particular attribute or not (see Fig. 2.3). Therefore, the annotations for the binary semantic attribute contain either 1 or 0.

2.1.1.2 Relative Attribute

The binary attribute-based representation encodes the presence or absence of certain attribute with a binary bit, which might restrict its expressive power. To address this issue, Parikh

Fig. 2.3: The example images with the binary attribute annotations for sport shoes, face, and chair, respectively.

et al. [105] proposed the novel concept of “relative attribute”, revealing the strength of the attribute with respect to other images. In contrast to the binary attribute with the description such as “this sport shoes is casual or not”, the annotation for the relative attribute is based on the comparison between image pairs like “this pair of shoes is more/less casual than that pair” (see Fig. 2.4). Such comparisons are much closer to human’s description of objects, which conveys richer semantic information. Kovashka et al. [74] integrated the relative attribute in the interactive human-provided feedback by specifying which properties of the image should be adjusted and this novel mode of feedback helped to improve the retrieval performance (e.g., shoes, faces, scenes). To further deal with the local attribute comparisons between similar examples, Yu et al. [157] proposed to learn the local-based ranking function to differentiate the subtle visual differences. Despite its resemblance to the way how human describe objects, the relative attributes suffer from one drawback that the attribute comparison annotations are difficult to collect and reach agreement in different annotators. The provided feedback in terms of the attribute strength might be biased from person to person, which might lead to an unreliable image search system.

Fig. 2.4: The binary attribute encodes the presence or absence of the “casual style” attribute with a bit. It is easy to define that (a) is casual while (c) is formal, not casual. However, in some cases it is difficult to draw a clear line about whether this shoe is casual or not, as shown in (b).

2.1.1.3 Multi-class Attribute

Another important type of semantic attribute utilized to describe the visual characteristics of the objects is the multi-class attribute, which can be considered as an organized and grouped form of the binary attributes. For example, the color attribute can take several values such as green, yellow, red, etc. Similarly, the category attribute of clothing can have multiple values such as blazer, T-shirt, browse, etc. In this thesis, the binary attributes are adopted in Chapter 3 to describe the properties of shoes and Chapter 4 as well as Chapter 5 handle the multi-class semantic attributes of shoes for visual representation.

2.1.1.4 Fine-grained Attribute

The earliest definition of attributes [41, 82] is supposed to be generic so that the concepts can be shared among different categories. In recent years, the fine-grained classification have attracted great attention among researchers, and it becomes essential to introduce more fine-grained attributes to describe the appearance and details of the part of the objects that distinguish different subcategories of the base category (see Fig. 2.5). The task of the fine-grained classification greatly benefits from the introduction of the fine-grained attribute into the local part, which captures the subtle visual differences between different subcategories of a basic category (e.g., bird [16], flower [7], aircraft [139]). Examples of fine-grained local attributes include

Fig. 2.5: Several examples for the local fine-grained semantic attributes for the dress, bird, and shoes. The localizations for the attributes are highlighted in the green and yellow bounding box for the associated attributes in the same font color.

the color at the wing, shape of the pedal, etc. Vedaldi et al. [139] investigated the influence of part detection and semantic attributes on understanding the fine-grained aircraft in detail. Berg and Belhumeur [16] proposed Part-based One-vs-One Features (POOFs) which extracted local features from the semantic parts of the objects for fine-grained classification and attribute estimation. Semantic attributes were also demonstrated to be effective in the fine-grained image retrieval problem. Li et al. [83] integrated the part-aware shoe attributes for fine-grained sketched-based image retrieval. Moreover, the study of visual clothing item has achieved great success attributing to the incorporation of fine-grained local attributes. Di et al. [147] presented a clothing dataset with rich annotations of fine-grained semantic attributes and the stylistic clothing attributes, which were further used for clothing retrieval and automatic tagging. Yamaguchi et al. [152] made use of the inter-object and inter-attribute compatibility to facilitate the detection of fashion items based on the street human image. Wang et al. [143] developed a clothing search system incorporating the BoW model for the initial ranking and exploiting the clothing attributes for re-ranking to improve the quality of retrieval results. Chen et al. [27] proposed a novel double-path based domain-adaption methods to bridge the gap between daily life and online store images to facilitate the subsequent application of human attribute description. As we know one of the drawback for the attribute-based representation is that the tedious

and error-prone procedure of annotation for the semantic attribute. To deal with the annotation issue, Berg et al. [14] proposed to automatically mine the useful attributes for fashion products from the noisy information provided by the web data.

2.1.2 Attribute-based Visual Representation

Fig. 2.6: (a) The concatenation of attribute prediction scores and (b) attribute prediction guided visual feature learning.

Existing works incorporating the attribute into the visual representation can be categorized into two directions: 1) the simple concatenation of the confidence scores or predicted binary/multi-class values from the attribute classifiers (see Fig. 2.6 (a)); 2) the low-level feature learning guided by the semantic attributes (see Fig. 2.6 (b)). The first is to use semantic attribute as a form of compact and discriminative feature for image representation, which is first proposed and employed in [41]. In the case of attribute representation based on binary attributes

composed of either 0 or 1, it is similar in spirit to the hamming vector. The key difference is that the former encodes the semantic meaning into the feature representation. Gong et al. [53] built a 3D object search system by matching the user provided descriptions to the annotated semantic attributes of database objects. Torresani et al. [136] employed the output of a set of classifiers learnt using semantic category information from the LSCOM ontology [101] for image categorization. Inspired by the success of [136], Douze et al. [38] further demonstrated that the performance of particular object retrieval benefited significantly from the combination of Fisher vector and the outputs from the attribute classifiers with the vocabulary from [136].

The alternative is to learn the visual representation for the objects guided by the semantic attributes, which have been successfully applied to a variety of applications, such as action recognition [89, 155], clothing recognition [26, 27, 93], person re-identification, instance-level image retrieval [24], etc. Chen et al. [24] combined the attribute-embedded features with the semantic-aware codes by incorporating the human attributes into the sparse coding framework in the task of content-based face image retrieval. Recent works began to incorporate attribute classification in the CNN learning process in order to achieve a better attribute prediction accuracy as well as to generate a more discriminative feature. Zhang et al. [172] developed a tree-structured poselet-specific CNN for human attribute prediction and further proposed an effective pose-normalization representation with features extracted from the poselet-specific CNN. However, their proposed CNN was used to predict the binary attributes. Huang et al. [63] integrated tree-structured CNN network with fine-grained clothing attribute labels when learning effective clothing features for retrieval. However, they merely utilized the attribute prediction as a regularizer for clothing retrieval. In these methods, the inherent correlation of the multi-class attribute is not taken into account. There are some existing works exploring the mutual attribute coherence to both improve the attribute prediction accuracy and develop more effective feature representation. Wang et al. [146] proposed to model the relationship between various attributes with a tree-structured graph. Chen et al. [26] developed an automated system

that was capable of describing the clothes with a list of pre-defined attributes for people on the street with unconstrained poses. It investigated the mutual dependencies between the attributes with a Conditional Random Field [81] to improve the attribute prediction accuracy. Nevertheless, they merely utilize the combination of hand-crafted features and the inherent properties of the multi-class attribute have not been fully explored. We further organize semantic attributes in a hierarchical manner as subordinate attributes in the CNN attribute classification framework for a more discriminative representation.

2.2 Instance-level Image Retrieval

The topic of content-based image retrieval (CBIR) with global descriptors (e.g., color [66], GIST [102], texture features [97]) has been extensively explored and the related techniques have reached some degree of maturity. In recent years, the instance-level image retrieval, a more challenging mode of searching for images based on the visual content of the query image, has gained increasing interests. Given a query image depicting a particular object, instance-level image retrieval aims at finding the images containing exactly the same object in the reference database. Recent works on instance-level image retrieval focus on either designing a robust image representation (feature-based) or developing the metric learning technique to evaluate the similarity between the query and reference images (metric learning based). We will review the previous studies in these two directions in the following and also present the image retrieval on the fashion domain, particularly on shoes. In the following, we use “instance-level image retrieval” and “instance retrieval” interchangeably. The typical instance retrieval framework is demonstrated in Fig. 2.7.

Fig. 2.7: The block diagram of a typical instance-level image retrieval system. Given a query image depicting a particular object (we take the shoes as a concrete example), instance-level image retrieval aims at finding the images containing exactly the same object in the reference database.

2.2.1 Feature Representation for Instance Retrieval

The research efforts of developing the visual representation for the instance retrieval problem can be mainly divided into two directions: conventional image representation and deep feature representation. The former utilizes the hand-crafted local image descriptors such as Histogram of Gradients (HOG) [31], Scale-invariant feature transform (SIFT) [95] aggregated into some encoding methods (e.g., Bag-of-Words (BoW) [132]). While the latter deals with the features extracted from the Convolutional Neural Network (CNN).

2.2.1.1 Conventional Image Representation

The conventional image representation employed in the instance retrieval focuses on describing the visual contents in the local aspects, and this line of approaches mainly consists

of three components: the interest point detector, local image descriptor, encoding technique. The interest point detector aims to localize the keypoints in the image which are stable to the local/global variation (e.g., scale, illumination changes) and representative image detectors include difference of Gaussians (DoG) [95] for the scale-invariant regions, the Maximally Stable Extremal Region (MSER) detector [98] for the affine-invariant locations, etc. As an alternative to the sparse-located keypoints, some researchers [65] adopted the densely-sampled image patches for feature extraction, which are expected to contain more detailed information of the image. Then the local descriptors (e.g., scale-invariant feature transform (SIFT) [95], speeded up robust features (SURF) [11]) are employed to describe the visual content of the detected regions or dense image patches, followed by the encoding step to embed the aggregated local descriptors into a compact feature representation with the codebook, which is learnt by clustering the local feature descriptors over the training images. The encoding schemes map the close local descriptors to similar dictionary words. Representative encoding schemes include Bag-of-Words (BoW) model [132], Fisher Vector (FV) [69], and Vector of Locally Aggregated Descriptors (VLAD) [9]. However, as reported by Tao [135], the local descriptor based methods lose their effectiveness for 3D generic objects like shoes due to the large viewpoint variation, which makes the descriptor matching less reliable.

2.2.1.2 Deep Feature Representation

These years have witnessed the emerging popularity of the deep learning based methods, which is capable of capturing the hierarchical representations of data with a cascade of layers, and a variety of deep architectures [55, 56, 62, 76, 110, 120] have been proposed. Among them, Convolutional Neural Network (CNN) [76] has been most widely used for processing visual data and achieved breakthrough performances in different areas such as image recognition [22, 37, 131], object detection [51, 112, 124], etc. Moreover, it has been shown effective to

Fig. 2.8: Example network architectures for AlexNet [75] and VGGNet [130]. Images from [161].

leverage the deep feature representation activated from CNN for image retrieval [10, 125, 162]. The performance with the deep features has a significant improvement over the conventional image retrieval pipeline using hand-crafted features like global features (e.g., GIST [102]) and local features with different encoding methods (e.g., DenseSIFT with Fisher encoding [69] or SURF with VLAD [9]). There are several models that are commonly adopted and have achieved significant improvements over conventional methods. Examples include AlexNet [75], VGGNet [130], etc. Fig. 2.8 demonstrates the overall architectures for these two representative convolutional neural networks. Before introducing the CNN-activated deep feature representation scheme, we firstly provide a brief warm-up review about most commonly used layers inside a CNN for feature extracting in the instance retrieval.

1) Convolutional Layers: The convolutional layers are the important building blocks for a typical CNN architecture. For each convolutional layer, it takes a 3D feature maps $z_i \in \mathcal{R}^{H \times W \times C}$ as the input, where H is the height, W is the width and D is the number of filters, with the first input as the raw image subtracted by the average image. Taking the AlexNet as a concrete example, it consists of five convolutional layers followed by three pooling layers and three fully connected layers, as shown in Fig. 2.9. An input image with the size of $224 \times 224 \times 224$ is fed forward into the CNN according to its specific structure. The convolutional layers consist of a set of kernels with small receptive fields and are utilized to compute the product

Fig. 2.9: The architecture of the AlexNet with Images from [76].

between the height and width of the image to capture the discriminative patterns of the input image. The shallow layers of CNN are supposed to capture the low-level cues such as the edge, corners and the deeper layers are supposed to capture more semantic information such as the parts of the object.

2) Pooling Layers: The pooling layers are located between neighboring convolutional layers, which aim to reduce the size of the feature map forwarded to the next convolutional layer. The pooling layers can also be considered as the sub-sampling layers which are to reduce the resolution of the features so that the processed features are more robust against noise and distortion. The common choices of the pooling layers include the average pooling and max-pooling layers, which are used to take the average value of neighboring regions in the feature map or return the maximum value of the local region, respectively.

3) Fully-Connected Layers and Loss Layer: The fully-connected layers are placed after a set of convolutional layer and pooling layers. These layers are basically the neurons in the previous layer connecting to all the neurons in the next neighboring layer. The image descriptor extracted from the fully-connected (FC) layer of the network (e.g., fc6, fc7 in AlexNet) is considered as a global feature representation. After appropriate normalization techniques such as L2-norm, L1-norm or power normalization [108] before computing the Euclidean distance between the query and reference images, the retrieval performance can be further improved.

The last layer of a typical CNN network is the loss layer, which differs according to the specific task on hand and penalizes the deviation between the true labels and the predicted values forwarded from its neighboring layer. The softmax loss is the most classical one in the CNN-based attribute prediction task, and the hinge loss is for the siamese network or the triplet network.

The deep features employed for the task of instance retrieval can be extracted from the pre-trained CNN model as well as the fine-tuned CNN structure. The former extracts the feature representation by forwarding the image into the pre-trained CNN in a single-pass. Here the pre-trained CNN is learnt with large collections of training images from ImageNet [118]. And the fine-tuned CNN is trained on the collections of images with the similar data distribution as the images in the target task [175]. The learnt feature is supposed to carry more domain-specific information.

1) From the pre-trained CNN: Jia et al. [37] demonstrated that CNNs pre-trained on the large-scale dataset, ImageNet [118], are good choices for feature extractors suited to the task of instance retrieval. Similar conclusion was raised by Razavian et al. [125], which demonstrated that the generic CNN feature was powerful for visual instance retrieval.

2) From the fine-tuned CNN: In [10], neural codes, a high-level descriptor invoked from the upper-layer of CNN, is capable of conveying the high-level abstraction of the image and demonstrated impressive performance. Following that, some researchers found that the local deep features activated from the lower-layers from CNN (i.e., convolutional layers) achieved better performances in instance retrieval tasks. Razavian et al. [111] suggested that features activated from the intermediate layers (convolutional layers) together with Spatial Pyramid Pooling (SPP) [60] can be used to better preserve the spatial information. Ng et al. [162] further compared the performance of deep features activated from different layers of CNN and found that VLAD encoded features from the conv5 layer outperformed other shallow layers in many instance retrieval benchmarks. Moreover, combining both features extracted from the fully-connected and

last convolutional layers could further enhance the retrieval performance [176]. The multi-scale CNN features are also beneficial for image representation. As reported by Gong et al. [54], they combined activations extracted from local image windows at multiple scale levels and achieved excellent results in scene retrieval.

Over these years, it has become increasingly popular to integrate semantic attributes into the feature learning process. Especially for the training of the multi-task CNN, the joint learning of several different losses can be used to better regularize the training of the network and lead to faster convergence speed, also less prone to overfitting problem. McLaughlin et al. [99] jointly optimized the task of the human attribute recognition and person identification to learn the deep semantic feature for the problem of person re-identification. Similar ideas have also been proposed in learning the attribute-based face representation [94] and in the fashion domain [93], etc. Liu et al. [93] proposed a FashionNet which learnt the clothing feature by simultaneously regressing the clothing landmark location and predicting the clothing attributes. It has been demonstrated that the activated feature embedding carries the semantic information and are more close to human perception of the objects. Different from these works that consider the attribute classification independently in the CNN learning, we design a novel definition of “shoe style”, which is a combination of different part-aware semantic shoe attributes, then the corresponding style identification loss is proposed.

However, there is few work to verify which layer is responsible for the improved performance for the cross-scenario retrieval framework. Inspired by [162], together with the reason that model-to-shop shoe retrieval falls into the field of instance retrieval, we also adopt the last convolutional feature as the feature representation for shoes in this work. Similar to the local descriptors in the conventional pipeline, there is a need to aggregate the CNN-based local descriptors into a global representation by encoding or direct pooling. Here it is worth noting that the same encoding is adopted for the CNN-based intermediate features with the popular

choices of VLAD or Fisher, which have achieved superior performances in Oxford building dataset [162]. Direct pooling over the feature maps such as the max pooling, sum pooling, and mean average pooling have achieved impressive performances compared with either convolutional layers or fully connected layers [176].

2.2.2 Metric Learning for Instance Retrieval

Besides a discriminative and robust feature representation, a proper similarity metric is also critical for an effective instance retrieval system. In the following, we review two different categories of the similarity metric: 1) pre-defined metric; 2) metric learning techniques. The first kind of metric leverages the existing distance functions to measure the similarity between different samples. Commonly used distance functions in the instance retrieval system include the Euclidean distance, cosine similarity, and hamming distance based on the compact image representation, etc. Euclidean distance and cosine similarity are the most widely used in the retrieval system with the low-level feature representation composed of real values, which computes the square root of the sum of the squared absolute differences and the dot product scaled by the product of their magnitudes of query image vector and the database image feature vectors. Another popular used metric is the hamming distance, which is employed in the case when the images are represented with several binary bits (e.g., 0 or 1) and the similarity between two images is the number of bits which differ between two binary strings.

The metric learning is intended to learn the transformation matrix that projects the raw feature to a low-dimension embedding such that data points of similar classes are drawn much closer simultaneously and those belonging to different classes are separated from each other. Representative metric learning algorithms include Large Margin Nearest Neighbor (LMNN) [148], Information-Theoretic Metric Learning (ITML) [33]. Both of these approaches are Mahalanobis metric based which learn a positive semi-definite matrix M so that the distance

$d_M(x_i, y_j) = (x_i - y_j)^T M (x_i - y_j)$ between the pairs of similar data points (x_i, y_j) is smaller than that of dissimilar pairs of data. LMNN penalizes the large distances between each input query and its target neighbors (K-nearest neighbors of the input query) to ensure they stay within the same class. Simultaneously, it discriminatively separates the data points belonging to different classes farther away. However, it is supposed to have the over-fitting problem during the training procedure for the lack of regularization [153]. ITML learns the distance function under the setting of information theory by maximizing the entropy of the multivariate Gaussian under the pairwise constraints [33]. These methods share the limitation of imposing the Positive Semi-Definite (PSD) or symmetry constraints on the learnt matrix M . However, Online Algorithm for Scalable Image Similarity (OASIS) algorithm explores the bilinear pairwise similarity function in the form of $S(x_i, y_j) = x_i^T M y_j$ between the pairs of data points. In practice, OASIS is quite fast and scalable to large-scale scenarios. We choose to use OASIS because it is the state-of-the-art approach [49] for learning the semantic similarity, which employs the large margin criterion for fast training convergence [23] without degrading the performance.

Recently, researchers have begun to integrate the metric learning into the framework of CNN, namely Siamese network and the triplet network. The former consists of two symmetric branches that take the matched pairs (same class) or non-matched pairs (different classes) as the input. It has been successfully applied to the task of product design [12], sketch-based image retrieval [140], image-patch matching [163], etc.

The triplet network is fed forward with three images as the input, corresponding to the anchor, the positive example with the same class label as the anchor and the negative example from a different class. The conventional triplet loss function [123] enforces the distance between the negative pairs (the anchor and the negative) larger than a pre-defined margin to the distance between the positive pairs (anchor and the positive). The goal is to pull images from the same class closer and push further away image pairs from different classes. The triplet network with three branches is reported to be superior to the siamese network [77].

Wang et al. [141] successfully built a multi-scale triplet network for learning the image similarity. The success of FaceNet [123] on face verification and recognition becomes the driven force that motivates the following work on the task of person re-identification [88], vehicle retrieval [87], unsupervised video representation learning [144]. The recent work by Zhang et al. [173] investigated the structure of the class label for fine-grained feature representation and Liu et al. [87] proposed a Coupled Cluster Loss (CCL) to further pull the positive vehicle images closer. Jiang et al. [70] proposed a scenario-based triplet weighting approach to address the clothing retrieval problem. Besides the success of triplet network on a variety of applications, researchers also raised some variants of the triplet loss functions to further handle different situations. A straight forward variant of the triplet loss function was proposed in [128], named ranking triplet loss. Rather than normalize each feature vector in the triplets, it enforces the pairs of feature distances to have the unit norm. Ripple et al. [115] proposed the *magnet loss* to encourage the samples from the same category to form clusters in the learnt embedding space. Sohn et al. [133] addressed the limitations of slow convergence of triplet network by incorporating more than one negative example in each update round. Hermans et al. [61] proposed *batch hard loss*, which is free of mining the hard negative examples offline.

The large-scale dataset makes the number of the candidate triplets grow exponentially. However, most of the triplets easily satisfying the triplet constraints produce close to zero gradient, which don't make any contribution to the parameter update. Thus, in order to make the training of the triplet network more efficiently, researchers proposed to mine the hard examples, including 1) hard negatives: anchor and negative examples are close; 2) hard positives: anchor and positive examples are far away. These two cases violate the triplet constraint and produce large gradients. Wang et al. [141] designed an online sampling scheme to select the hard example triplets based on the relevance score and the hard negative examples were drawn from the same distribution as that of the positives. And the difference of their relevance score with the anchor sample should be larger than a threshold. However, the feasibility and effectiveness of

this method is largely limited by the manual annotation. Serra et al. [129] proposed to sample the triplets based on the image labels. The training pairs for the hard positives were formed by sorting the randomly sampled pairs of anchors and positive images in descending order while the hard negatives were picked from the randomly sampled pairs of anchors and negative examples in ascending order. To address the poor training problem and the mislabeled faces in the dataset, Schroff et al. [123] presented online mini-batch triplet sampling scheme, which computed the argmax and argmin distance over the mini-batch in the training set. Moreover, he put forward the novel concept of “semi-hard negative”, the distance of which was close to the anchor positive distance, which could efficiently tackle the bad local minima problem of merely using the hard negatives. However, these methods only compute the feature-level distance to differentiate the hard examples and don't fully exploit the semantic information. In Chapter 5, the hard examples are evaluated in both the feature-level and attribute level. Furthermore, we develop an attribute-based hard negative weighting and sampling strategy for network training.

2.2.3 Image Retrieval for Fashion Items

Existing works on fashion items mainly target on clothing, like clothing attribute prediction [26, 93], clothes parsing [150, 151], clothing retrieval [59, 91, 93, 100, 178], etc. He et al. [59] developed a mobile product search system utilizing the bag of hash bits for generating initial results and boundary features for reranking. Mizuochi et al. [100] presented a clothing retrieval system which enabled the user to specify the portion of the image they don't like and considered the local similarity of multiple images to return a desired image. Among these works on the fashion image retrieval, both the query and the reference images are from the same domain, online shops with white background and pictured in ideal environment. Moreover, there exists several works addressing the domain-gap in the clothing retrieval problem. Liu et al. [93]

developed a multi-task CNN that jointly predicted the attribute and landmark location and the features activated from the learnt CNN were used to perform the clothing retrieval. Chen et al. [27] developed a deep domain adaption network to model the clothing features extracted from the different domains simultaneously keeping the consistency of matched pairs between different domains. Zhou et al. [178] proposed a hybrid topic model to bridge the gap between semantic text descriptions and extracted low-level features.

In contrast to the clothing retrieval, the studies on image-based shoe retrieval is still at infancy stage and only few works are reported. Kovashka et al. [74] developed a shoe recommendation system utilizing relative attribute based pairwise comparison. However, it is difficult to get a consensus on the relevance feedbacks (e.g., shoe A is feminine than shoe B) among annotators. Huang et al. [64] proposed a shoe retrieval framework via part-related attributes, which jointly optimized the tasks of Deformable Part Model (DPM) based shoe part detection and attribute prediction, which is more relevant to our work presented in Chapter 3. Nevertheless, it suffered the slow detection speed due to the use of DPM. This method employs a set of small bounding boxes to represent each part, which may not contain sufficient discriminative information. Moreover, the query image captured from the single viewpoint is not enough to depict the appearances of the shoes. Kiapour et al. [57] focused on the exact street-to-shop product retrieval by adopting a category-independent metric network to evaluate the similarity of image pairs. For shoe image retrieval, commonly the location of the query object is provided as a prior knowledge. However, in our work, a region proposal selection approach to localize the shoes is proposed and the top-3 selected region proposals are regarded as one level of the feature of our system. We represent shoe features using a three-level structure combined with semantic hierarchy of attribute-based CNN, which facilitates us to learn the semantic-aware features for cross-domain exactly the same shoe retrieval. We also focus on the hierarchical properties of the semantic shoe attributes, and propose the subordinate attribute to improve the prediction based on these discriminative attributes.

2.3 Multi-Task Learning

Multi-Task Learning (MTL) jointly optimizes several different but related tasks to achieve better generalization performance, compared with solving multiple tasks independently. Especially when the labeled training samples are limited, the single-task learning strategy might result in high generalization errors. Under the setting of multi-task learning, the information contained in the training signal for each task is considered as the domain-specific inductive bias for the rest tasks [21] and the shared information between different tasks is explored. More specifically, as all the tasks are noisy to some extent and have distinctive noise patterns, multi-task learning learns a shared and more general representation for all tasks by averaging those noise patterns. But there still remains one question: why the multi-task learning is effective than the single task? Ruder in [117] summarized the reasons into four points: 1) implicit data augmentation; 2) attention focusing; 3) eavesdropping; 4) representation bias; 5) regularization.

Its effectiveness has been proven in a variety of computer vision areas such as face landmark localization [174], face detection [170], pedestrian pose estimation [52], etc. The feature learning process also benefits from exploiting the intrinsic relationships (differences and commonalities) between different tasks [36, 72, 84]. Nevertheless, it is still new in the fashion area, especially on shoes, partly due to the absence of accurate attribute labels. Because the descriptions of fashion images crawled from the online website are not reliable. The most relevant work is [93], which learns the clothes feature by jointly optimizing the clothing landmark detection and attribute prediction. The work presented in Chapter 5 combines the multi-task learning with the triplet loss network for learning a more semantic feature representation for shoes. And the attribute annotations are provided in our newly-built dataset. Our work differs from [93]

in the sense that the retrieval of shoes has its unique viewpoint variation challenge. To address this problem, we design a viewpoint invariant loss over the multi-task framework and obtain a view-invariant shoe feature representation.

2.4 Object Detection

Object detection can be considered as the process of localizing the positions and the scales of the object instance via the rectangle bounding box. This can be viewed as a pre-processing step for the subsequent image retrieval. With the main focus on the localized target object, it could benefit the efficient searching and retrieval.

In the past decades, the approach in [31] utilizing the sliding window templates with the discriminative HOG features has dominated the area of object detection for a long time, the core idea of which is to scan the entire image with candidate windows in various scale, positions and aspect ratios and to search for the regions indicating the target object. However, it suffers from the large computation burden due to the fact that it needs to search exhaustively on the image to localize the possible regions of interest. Thereafter, the emerging of the deformable part model (DPM) [45] composed of a coarse root HOG filter for the entire object and several higher resolution part filters achieved the state-of-the-art performances on several benchmark datasets (e.g., PASCAL VOC [40], INRIA Person dataset [31]) with highly-varied object classes and has become the leading technique in the computer vision community. It leveraged the mixture of components to model the variations in pose and viewpoint of the objects and was experimentally demonstrated to be effective in handling those object category with the large intra-class variance (e.g., pedestrians and cars). For example, Zhang [171] trained a DPM model for pedestrians with three components roughly corresponding to the front, side and three-quarter views. Moreover, this method is able to localize discriminative object parts without rich

annotations of the part location. This is especially helpful for those objects with explicit part structure such as shoes, cars, etc. The star-cascade version of the DPM [44] has successfully dealt with the speed bottleneck of the original DPM and made it applicable to real-time detection scenarios.

The object region proposal based method is designed to address the trade-off between the computation tractability and satisfying detection performance. By analyzing the low-level cues, detection proposals aim to rapidly determine the candidate windows with high probability including the target object instances. Representative region proposal based methods include selective search [137], EdgeBox [179], etc. Selective search [137] greedily merged the superpixels in multiple hierarchical segmentations to determine the final region proposals. EdgeBox [179] considers the number of contours as an important cues to generate the region proposals. Based on the generated proposals, the following steps to localize the object can be classified into two categories: classifier-based method and the ranking based method, as demonstrated in Fig. 2.10.

The typical example of classifier-based method is the R-CNN framework [51] and its variants [50, 114]. The pipeline of R-CNN [51] was to feed the initial large pool of region proposals into the network for feature extraction, then build category-independent classifiers to localize the object based on the likelihood of candidate bounding box belonging to that category. Afterwards, Fast R-CNN [50] proposed to improve detection efficiency by incorporating the region proposal map and Faster R-CNN [114] proposed the Region Proposal Networks (RPN) for faster detection speed.

The ranking-based method aims to develop a ranking function to re-rank the initially generated region proposals based on the bottom-up visual cues (e.g., saliency and edge) and choose the top-K candidates as the final list. The recent state-of-the-art detection approach DeepBox [80] pruned the false positives candidate boxes via specially-designed CNN, and achieved outstanding performances. Zhong et al. [177] extracted the low-level and semantic cues (e.g.

Fig. 2.10: The pipeline of object detection framework utilizing the region-proposal based object detection utilizing the ranking-based methods and classifier-based methods.

semantic segmentation, contextual information) and re-rank them based on the score returned by structured SVM for the application of automatic driving.

The above works essentially utilized two-stage detectors by firstly generating a sparse pool of candidate object locations and then classifying or ranking them as one of the foreground categories [85]. Recently, another line of work [47, 92, 112, 113] proposed to perform the object detection with faster speed and is free of generating region proposals. You Only Look Once (YOLO) [112] is an extremely fast real-time detector, which utilizes a single deep neural network to predict the bounding box locations and their class probabilities from the full image, without producing any region proposals. To further improve the speed and detection accuracy, Redmon et al. [113] introduced improved versions of YOLO, named YOLOv2 and YOLO9000, to detect a wider range of object categories at varying image sizes. With the labeled classification data from ImageNet [34] and the detection data from COCO dataset [86], YOLO9000 is capable of detecting more than 9000 categories of objects by jointly optimizing the classification and detection tasks. The Single Shot Detection (SSD) [92], also detects the object locations in a single-pass and is capable of dealing with objects in varying sizes and scales by developing a pyramid of feature maps with different resolutions. To address the problem of detecting small objects, Deconvolutional SSD (DSSD) [47] introduced deconvolutional layers to incorporate large-scale contextual information.

2.5 Summary and discussions

In this chapter, the attribute-based representation, instance retrieval, multi-task learning as well as the object detection techniques are introduced. Over the past decades, the computer vision researchers have devoted considerable efforts into developing effective methods in the above mentioned research domains. The conventional methods are gradually replaced by the CNN-based methods and feature representation due to their superior power, which verifies the necessity of incorporating the CNN networks. The shoe is a special category of the fashion item with its own unique challenges, such as the large viewpoint variation, which makes the local descriptors matching ineffective. In addition, the semantic attribute based representation can be further improved based on the intrinsic relationships of the multi-class attributes, which leads to an improvement in the the retrieval accuracy. On the other hand, most of the existing instance retrieval systems are merely based on one query image, which might not provide enough information for the shoes. Even in some viewpoints, the appearance of shoes is partially visible. Therefore in Chapter 3, we design a novel attribute-based shoe image indexing and in-store shoe retrieval framework with the multi-view images, in which we make use of the complementary information provided by multiple viewpoints. Moreover, the semantic attribute values are jointly determined by several viewpoints under which the attribute is visible. The system is simple yet effective, and it shows an enhanced performance in attribute prediction and shoe retrieval. Compared to the semantic attribute based representation that merely use the output of the attribute classifiers as the final output, our representation is specially designed for shoes.

CNN has achieved superior performances for different kinds of vision-based tasks, which motivates us to incorporate the CNN-based method in the shoe attribute prediction and retrieval. Previous CNN models on the attribute prediction don't explore the intrinsic relationship between the multi-class semantic attributes, especially on shoes. Moreover, they put equal penalty

on the prediction results for different attribute which is not fair. Surely, the error prediction of some attributes should be punished more severely than the rest. Motivated by this, in Chapter 4, we consider the attribute hierarchy and design a multi-level deep shoe feature representation to overcome the challenges raised from the cross-domain shoe retrieval.

Even though the triplet network has gained great popularity in the retrieval problems such as vehicle retrieval, person re-identification. There are not many researchers specially applying it to the product retrieval. In Chapter 5, we are still doing the topic of cross-domain shoe retrieval while the images from the online stores are in a variety of viewpoints. In this chapter, our network is built based on the triplet network and some constraints are incorporated to consider specifically on the shoes problem with viewpoint variation and fine-grained differences. Furthermore, we design the hard example choosing selection strategy based on the consistency between the feature level and the semantic-space level. Finally, we build a multi-view shoe dataset to verify the effectiveness of our approach, and the results indicate that we can return shoes of multiple viewpoints as the input.

Chapter 3

Towards Multi-View Shoe Image Indexing and Retrieval

In Chapter 3, we propose a novel shoe image indexing (i.e., automatic shoe attribute prediction system) and retrieval system¹ based on the multi-view shoe images, where each shoe is indexed by a list of part-aware shoe attributes. Firstly, the viewpoints of input images are identified. To estimate the value of each shoe attribute, we propose to select only a set of relevant view images according to the shoe part which accommodates the attribute. Then, based on the prior knowledge of the shoe structure under a certain viewpoint, a view-specific part localization model is proposed to localize the shoe part from each of the relevant views. Proper features are then extracted from all the localized shoe parts for estimating the attribute value. We measure the similarity between different shoes based on the estimated index, which is used for the final retrieval. The effectiveness of the proposed system is demonstrated on two newly-built shoe datasets.

¹Part of this chapter was presented in [166].

3.1 Introduction

The problem of visual object retrieval has been studied over decades and a large body of algorithms have been developed on a variety of object categories, such as Oxford buildings [10], objects in MSR-Bing dataset [121], annually updated dataset of TrecVid Instance Search [103], INRIA holidays dataset [68], etc. Relatively few work focus on the visual product retrieval, especially on the fashion items, which is in great need with the increasing popularity of online shopping. In many cases, it is difficult for the consumer to search for fashion items (e.g., clothes, bags or shoes) by using key words. For example, the consumer may want to search for similar dresses she saw on a celebrity or her friends. To make people's life easy when searching fashion items, image based fashion item retrieval is necessary to be developed. On the other hand, for those online merchants selling fashion items, a great many new products are uploaded every day, it is also necessary to develop effective automatic indexing techniques. This would offer great help to manage these products as well as to facilitate the product search.

The traditional image retrieval techniques extract low-level features to analyze the content of an image [32] which is represented by a visual-word dictionary learned from clusters of local image patches [95]. To enable efficient search, inverted indexing [25] is usually employed to index each image in the dataset. The learned dictionary could capture the discriminative patterns of the image, however, it fails to model the higher level concepts of the image in a semantic way [159]. To tackle this issue, a mid-level concept "attribute" has been proposed in recent years to address a variety of visual recognition tasks [79, 106, 119] as well as image retrieval [122, 127]. It could be detected by machines and are also meaningful to human, which bridges the interpretation gap between machine and human.

The attribute based image representation, which is also know as semantic indexing, is shown to be a proper way for fashion item retrieval. One of the main reasons is that such techniques allow the consumer to perform a concept based search besides the image based retrieval. For

example, without the input of an image, the consumer can also search for shoes of “high heel” or dress of “A-line” using keywords. In recent years, attribute-based retrieval techniques have been developed on the search of fashion items including clothes [18, 35, 90], bags [145] and shoes [64, 74].

Among the works for the fashion item search, the studies on image based shoe retrieval is still at its infancy stage. The work in [64] proposes a shoe retrieval framework via part-related attributes. It jointly optimizes shoe part detection and attribute prediction based on Deformable Part Model (DPM) [43]. Localizing the shoe part is time consuming due to the use of DPM. This method employs a set of small bounding boxes to represent each part, which may not contain sufficient discriminative information. All these works consider to retrieve shoes based on the input of a single view shoe image which is usually not enough to have a good description of the shoes.

In this chapter, we focus on the semantic indexing and retrieval of shoes with a set of multi-view shoe images as input. This is motivated by the observation that majority of the online shoe stores display their shoes with a set of images in common views such as frontal, back, side and etc. On the other hand, social network users also have the tendency to post the images of their beloved shoes in those similar views. With such multi-view information available, the indexing and retrieval of shoes could be more accurate.

Concretely, we propose a novel shoe image indexing and retrieval system based on the multi-view aspects of the shoe information. Each index consists of seven binary element corresponding to seven part-aware shoe attributes. Given a set of multi-view shoe images, we first identify the view of each image into one of the six common shoe views which are predefined. For each shoe attribute (termed as attribute for short), our system finds a set of relevant views and then estimate its value based on these views. During the attribute estimation, a novel view-specific part localization method is proposed to localize the shoe parts (which accommodates

the attribute) on the relevant views. Subsequently, the HOG features are extracted from the localized parts for attribute estimation. Eventually, the estimated index of the query shoe with multi-view information is compared against those indexes in the dataset to find similar shoes. We conduct the experiments on two newly built shoe databases, one consists of 35670 multi-view shoe images collected from online merchants, the other contains 930 daily shoe images collected from social websites. The experimental results on the shoe attribute estimation improves over 4% in accuracy when compared with the state-of-the-art system. We also show the advantage of using multi-view shoe images for indexing and retrieval.

To the best of our knowledge, the proposed shoe indexing and retrieval system is the first one that is specially designed to deal with the multi-view query input. Moreover, a simple but effective view-specific shoe part localization technique is developed to localize the shoe parts assuming that their part appearances are pretty much structured for a certain view. A novel online shoe image dataset, composed of 5945 shoe items with 7 types of fine-grained shape-based attributes and annotated part locations, is collected to verify the effectiveness of our shoe indexing and retrieval system.

The rest of the chapter is organized as follows. In section 3.2, we explain how the multi-view shoe dataset is built as well as the corresponding annotation process. Section 3.3 introduces our proposed multi-view shoe indexing and retrieval system in detail. In section 3.4, the experimental results are shown with some discussions. Finally, the work is summarized and concluded in section 3.5.

Fig. 3.1: Examples of common viewpoints of a shoe and the manually annotated parts for shoe images in our dataset (best viewed in pdf file and color images).

3.2 Multi-View Shoe Dataset

3.2.1 Multi-view Shoe Images Collection

Currently, there is only one public available shoe dataset [14] which contains merely a single side view for each shoe. We find that most of the online shoe merchants (such as 6PM.com, Amazon.com, Zappos.com, etc) prefer to display the shoes in several multi-view images. We collect the multi-view shoe images from Amazon.com. These shoe images fall into seven common views, i.e., left-profile, frontal, back, bottom, left facing, right facing and top, as shown in Fig. 3.1 (a). For simplicity, these views are termed as $V_1, V_2, \dots, V_7$. The bottom view (i.e., view V_4) mainly shows the bottom part of the shoes, which is not very useful for presenting the appearance of the shoes. Therefore, it will not be included in our dataset. In total, our multi-view shoe dataset consists of 5945 women shoes (including pumps, flats and sandals), while each shoe is demonstrated around six views (i.e., there are $5945 \times 6 = 35670$ shoe images). In the following discussion, these views are denoted as $\mathbf{V}$, where $\mathbf{V} = \{V_1, V_2, V_3, V_5, V_6, V_7\}$.

Besides, we collect a daily shoe dataset, which consists of 930 shoe images in un-constrained

Table 3.1: Part-aware attributes and the corresponding relevant views.

Shoe Parts	Related Attributes	Relevant Viewpoints
head	closed toe, pointy	$\{V1, V2, V5, V6, V7\}$
back	back-covered	$\{V1, V2, V3, V5, V6, V7\}$
body	side-covered, bounds	$\{V1, V2, V5, V6, V7\}$
heel	high-thin, wedge heel	$\{V1, V3, V5, V6\}$

Fig. 3.2: Illustrations of the appearance for each shoe attribute. The first row of images correspond to the presence of each attribute, while the second row indicates the absence of the attributes.

poses, crawled from social websites like flickr.com. The shoes are presented under un-constrained poses and example shoe images are illustrated in Fig. 3.1 (b).

We define the index of each shoe as a combination of seven binary elements corresponding to seven part-aware attributes, including closed toe, pointy, back-covered, side-covered, bounds, high-thin heel and wedge heel. Each attribute is defined based on one of four shoe parts including head, back, body and heel. The relationship between the shoe attributes and shoe parts are given in Table 3.1. For better illustration, Fig. 3.2 provides several examples regarding the presence and absence of each attribute.

3.2.2 Annotation of shoe part locations and attributes

We asked 10 persons to annotate the shoe part locations as well as the attributes for each shoe. Given a shoe image, we annotate the locations of the shoe parts as bounding boxes. Each bounding box is required to cover the corresponding shoe part and represented as the

Table 3.2: The visible parts of each viewpoint.

Viewpoint	Visible Parts
V_1	Shoe head, back, body, heel
V_2	Shoe head, back, body
V_3	Shoe back, heel
V_5	Shoe head, back, body, heel
V_6	Shoe head, back, body, heel
V_7	Shoe head, back, body

coordinates in pixels of two corners (i.e., the top-left corner and the bottom-right corner). For each part, the ground truth location is computed as the mean average coordinate for the 10 bounding boxes annotated by the volunteers. It should be noted that not all the views visually contain the four aforementioned shoe parts. Table 3.2 lists the shoe parts that are visible under each of the six views in $\mathbf{V}$, and Fig. 3.1 also illustrates the ground truth shoe part locations for shoe images from two datasets. As we have mentioned before, the view V_4 is not included in the dataset thus we won't consider it in the part localization process. For attribute annotation, similarly, we mark whether an attribute is presented with value 1 or 0. The ground truth value of each attribute is determined by the majority vote among the 10 persons. To the best of our knowledge, this is the first shoe dataset which consists of structured multi-view images with annotations of shoe part locations and attributes.

3.3 The Proposed System

3.3.1 System Overview

Fig. 3.3 illustrates the flowchart of our proposed system for multi-view shoe image indexing and retrieval. Given a set of M ($M \geq 1$) multi-view shoe images S_m ($m = 1, 2, \dots, M$), we first identify the view of each image into one of the six views we defined earlier. Then, we select a set of relevant views for each attribute based on the corresponding shoe part, and feed them into

Fig. 3.3: The proposed retrieval system with multi-view query shoe images.

an attribute estimator to estimate the value of the corresponding index element. The attribute estimator localizes the shoe part that accommodates the attribute for each of the relevant views. This is done by using our proposed view-specific part localization model. After that, we extract the HOG features from the localized part and the ultimate value of the index element is based on the majority vote decision of those predictions from the relevant views. Finally, those shoes with similar indexes in the dataset will be retrieved based on the proposed index distance measure.

3.3.2 View Identification

We identify the view of a shoe image S_m into one of the six views in $\mathbf{V}$. Specifically, we build a view identification classifier based on the GIST [126] feature using a set of training data. The same feature is extracted from S_m and fed into the classifier to estimate the view.

3.3.3 Relevant View Selection

Let's denote the identified view of image S_m as V_m , where $V_m \in \mathbf{V}$. As what have been shown in Table 3.1, some shoe parts are not visually available under certain views. For each attribute which is part-aware, it is necessary to select only a set of relevant views (from the identified views) to estimate its value. That is to say, to estimate the value of a part-aware attribute, we need to find out the relevant views which visually contain the corresponding shoe part. Such relevant view selection is predefined according to the visual cues of different views in $\mathbf{V}$, given in Table 3.1.

3.3.4 View-specific Part Localization

In this section, we propose a novel shoe part localization method which is view-specific. Our proposal is based on the observation that the shoe images are pretty much structured for a certain view. For example, for the left-profile view V_1 (see Fig. 3.1 (a)), the shoe head is located on the bottom-left of the shoe region, while the shoe body is located around the center. This motivates us to develop a simple but effective shoe part localization model elaborated below.

First of all, the deformable part model (DPM) [43] is employed to estimate the foreground shoe region, which is known for its capability of detecting objects in cluttered images even with great variations in shape, appearance and camera viewpoint. The objects are represented using a coarse-scale root template and several higher-resolution part templates, with the geometric constraints on the locations of the parts. By applying the DPM model on the shoe images, the irrelevant background objects could be removed and we could mainly focus on the foreground shoe region. For the query images with clean white background, we implemented a simple image processing technique to remove the white spaces around the specified shoe region by counting the number of near-to-white pixels, T (i.e., the pixel value is larger than 240). When

processing with a row or column of image, if T is larger than a threshold, then we will remove the corresponding row/column from the image; otherwise it is remained. In the following discussions, a shoe image is referred to as the one after such background removal.

Let $I = \{I_i^a\}_{i=1}^N$ denote N shoes for training the part localization, where $a = \{1, 2, 3, 5, 6, 7\}$ indicates one of the views in $\mathbf{V}$. For example, I_i^a refer to the i th training image under view V_a . For each shoe image I_i^a , we define its annotated part locations (in terms of bounding boxes) as $B_{ip}^a = [X_{ip}^a(1), X_{ip}^a(2), Y_{ip}^a(1), Y_{ip}^a(2)]$ of the bounding box, where $p = \{1, 2, 3, 4\}$ refers to one of the four shoe parts, where $X_{ip}^a(1), X_{ip}^a(2)$ and $Y_{ip}^a(1), Y_{ip}^a(2)$ indicate two horizontal and two vertical coordinates of the bounding box respectively. We obtain the following part localization model for each view V_a by

$$\begin{cases} \overline{X_p^a(k)} = \frac{1}{N} \sum_{i=1}^N \frac{X_{ip}^a(k)}{W_i}, & k = 1, 2 \\ \overline{Y_p^a(k)} = \frac{1}{N} \sum_{i=1}^N \frac{Y_{ip}^a(k)}{H_i}, & k = 1, 2 \end{cases} \quad (3.1)$$

where W_i and H_i represents the width and the height of the training image I_i^a .

Given an input shoe image S_m with the view $V(m) = V_a$ identified, its shoe parts can be localized as $B_{tp} = [X_{tp}(1), X_{tp}(2), Y_{tp}(1), Y_{tp}(2)]$ which can be calculated below:

$$\begin{cases} \overline{X_{tp}(k)} = W_t \cdot \overline{X_p^a(k)}, & k = 1, 2 \\ \overline{Y_{tp}(k)} = H_t \cdot \overline{Y_p^a(k)}, & k = 1, 2 \end{cases} \quad (3.2)$$

where W_t and H_t records the width and the height of S_m .

3.3.5 Feature Extraction and Classification

For each attribute, we extract the HOG feature on the localized parts from the relevant views. As the attributes defined are related with the shape aspects of shoes, the boundary of the shoe could provide more visual cues in contrast with the surface regions of the shoes. To have a better representation of the shoe boundary, besides densely extracting the HOG features, we also extract these features from Harris corners [58] of the localized part. The overall features are a concatenation of both features. Due to the attribute we define are binary with values equal to 0 or 1, we train a set of binary SVM classifiers after the features are fed into the classifier. With these features, we train a set of binary SVM classifiers for each attribute mentioned above. And the final attribute value is determined based on the majority vote.

3.3.6 Index distance measure

Once the value of each attribute is estimated, the index of the input shoe is then set as the concatenation of all the estimated attribute values which are binary. It should be noted that, for some extreme cases, we are not able to perform an estimation for all the attributes. For example, if we only have $M = 1$ input shoe image which is identified as the view V_2 , the attribute estimation can only be done for “closed toe, pointy, side-covered and bounds” as shown in Table 3.1. Most of the time, all the seven predefined attributes can be properly estimated given an input shoe with multi-view images. Without the loss of generality, we denote the index of a query shoe as h_q , where

$$h_q(i) = \begin{cases} 1 & \text{if the attribute is present} \\ 0 & \text{if the attribute is absent} \\ -1 & \text{if the attribute is not in view} \end{cases} \quad (3.3)$$

and $i = 1, 2, 3, \dots, 7$ refers to the seven attributes. We measure the distance between h_q and the one in the dataset, say h_s , as below:

$$d = \sum_{i=1}^7 |h_q(i) - h_s(i)| \Gamma(i) \quad (3.4)$$

where

$$\Gamma(i) = \begin{cases} 1 & \text{if } idx_q(i) \neq -1 \text{ and } idx_s(i) \neq -1 \\ 0 & \text{Otherwise} \end{cases} \quad (3.5)$$

Here $\Gamma(i)$ is a indicator referring to whether the attribute i can be estimated from both the query shoe and the one in the dataset. The retrieval is done based on the distance d . The smaller distance d is, the more similar it is between the shoe in the dataset and the query.

3.4 Experimental Results and Discussions

We partition the 5945 shoes in our multi-view shoe dataset into three non-overlapped parts:

1) 1500 shoes are used in the training phase of several modules including the viewpoint identification, view-specific part localization and attribute prediction; 2) 4000 shoes are considered to be stored in the dataset for retrieval, which is termed as the 4000 shoe dataset in the following discussions and 3) the rest 445 shoes are treated as the queries for evaluating the shoe retrieval performance. Five-folder cross-validation are performed on the training set to determine the parameters of the SVM classifier. It should be noted that to validate the effectiveness of our proposed system, experiments are carried using 930 images from the daily shoe dataset and the accuracy of attribute estimation is demonstrated in the following section.

Table 3.3: Shoe part localization accuracy (in terms of PCP metrics) using our proposed method (shown in bold number) and the work in [64] (shown in non-bold number).

	Shoe Head	Shoe Side	Shoe Heel	Shoe Back
V_1 View	71.2\83.3	73.8\84	75.3\88.6	70.6\94.8
V_2 View	86.5	92.5	N.A	96.0
V_3 View	N.A	N.A	98.5	98.0
V_5 View	73.4\89.5	78.2\84.6	72.9\92.1	80.2\98.5
V_6 View	74.2\88.7	76.3\85.3	75.1\90.4	82.1\97.6
V_7 View	86.2	81.3	N.A	92.3

3.4.1 Performance of Viewpoint Identification

We apply the default settings for the GIST feature with the dimension equal to 3200 and learn the classifier with linear kernel SVM [13]. The view identification accuracy is 98.3% when tested on the 4000 shoes in the dataset. From the failure cases, we can find that the view identification classifier makes confusion at classifying the V_1 view and V_5 view, especially the shoes presented in V_1 is almost in the left facing view. Meanwhile, if the shoes are displayed in a confusing viewpoint, it poses difficulties for human to annotate the corresponding viewpoint labels. Moreover, the V_2 view is easily mixed with the V_3 view for flat shoes without much texture and heel.

3.4.2 Performance of Attribute Estimator

The key component of our attribute estimator is the view-specific part localization. In this section, we will evaluate the performance of the proposed view-specific part localization, followed by the evaluation of attribute estimation. Comparisons with the state-of-the-art work [64] will also be conducted. Note that, in [64], the input is a single-view shoe image belonging to the left-profile (V_1), left-facing viewpoint (V_5) or the right-facing viewpoint (V_6).

Evaluation of View-specific Part Localization: The accuracy of part localization is evaluated based on the area overlap between the ground truth part location and the detected one. The

Table 3.4: Attribute prediction accuracy of our proposed method (shown in bold number) and the work in [64] (shown in non-bold number) for different views.

Attribute type	V_1 View	V_2 View	V_3 View	V_5 View	V_6 View	V_7 View
Closed Toe	80.27\86.0	86.1	<i>N.A</i>	83.21\87.6	80.2\86.8	87.2
Back Cover	94.6\98.3	97.1	98.0	95.7\98.6	94.5\97.3	97.5
Side Cover	83.9\93.1	90.1	<i>N.A</i>	82.4\91.5	82.8\91.2	90.2
Pointy	87.2\92.2	91.4	<i>N.A</i>	88.2\92.7	87.6\92.1	93.1
High Thin Heel	80.2\89	<i>N.A</i>	87.3	83.6\90.2	82.8\90.4	<i>N.A</i>
Wedge Heel	86.7\95.0	<i>N.A</i>	93.4	89.4\96.1	87.3\95.8	<i>N.A</i>
Bounds	79.8\86.1	82.3	<i>N.A</i>	80.6\87.1	80.2\86.5	85.4

classical localization evaluation metric PCP (Percentage of Correctly Localized Parts) [154] is adopted here. The localization is considered to be correct if the detected part has more than half overlap with the groundtruth part. The result of the view-specific part localization on the 4000 shoe dataset is given in Table 3.3, where *N.A* indicates those parts that are not visible under the specific viewpoint. The work in [64] builds the part localization models for view V_1 , V_5 and V_6 viewpoints, the results of which is shown in italic number in Table 3.3. We can observe that, by using our proposed method, the part localization of Shoe Back achieves the best performance (over 92% in accuracy). Moreover, it outperforms the state-of-the-art work with more than 10% in accuracy.

To demonstrate the effectiveness of our proposed view-specific part localization model, we present the detected parts of several exemplar shoes with single-view and multi-view inputs as shown in Figure 3.4. It can be seen that our model is capable of predicting consistent part locations for different types of shoes, e.g., with straps, heeled shoes with different heights, wedges, etc. Even though the bounding boxes include the un-necessary white background, they still capture the most informative parts of the images.

Evaluation of Attribute Estimation: The effectiveness of attribute estimation performance is analyzed based on per attribute estimation mean average accuracy. We first compare the attribute estimation accuracy achieved by our work with the one proposed in [64] for each of the views in the 4000 shoe dataset, the results of which is given in Table 3.4. As mentioned

Fig. 3.4: Part localization results of the proposed view-specific part localization approach. In the result, the bounding box in different colors represent the detected shoe parts. (a) single view shoe images. (b) five viewpoint images for heeled shoes with different heights.

Table 3.5: The performance of attribute estimation on the “DailyShoes” dataset using the proposed method and the state-of-the-art [64].

	Proposed	State-of-the-art [64]
Closed Toe	78.3	71.6
Back Cover	90.5	84.1
Side Cover	86.2	78.1
Pointy	82.4	79.8
High Thin Heel	82.7	78.1
Wedge Heel	91.3	87.2
Bounds	88.2	84.3

before, we only conduct the experiments on view V_1 , V_5 and V_6 for the work in [64]. We can see from Table 3.4 that our proposed method outperforms the work in [64] regardless of the attributes, where the average improvement is around 5% in estimation accuracy.

Table 3.5 illustrates the attribute estimation result on the “DailyShoes” dataset using our proposed method and the state-of-the-art method [64]. It can be seen that our method still works better than the state-of-the-art with around 4% improvement in estimation accuracy.

Next, for each shoe from the 4000 shoe dataset, we choose $M = 1, 3$ and 6 views. Then,

Table 3.6: The performance of attribute estimation for a set of M views of each shoe using our proposed method (shown in bold number) and the comparison with [64] when $M = 3$ (shown in non-bold number).

	Single-View ($M=1$)	Multi-Views ($M=3$)	Multi-Views ($M=6$)
Closed Toe	86.3	87.4 \ 81.3	88.2
Back Cover	96.2	98.5 \ 95.0	98.8
Side Cover	92.5	93.7 \ 82.6	93.0
Pointy	92.0	92.8 \ 87.5	93.8
High Thin Heel	89.6	90.0 \ 82.4	91.2
Wedge Heel	94.1	95.6 \ 87.5	95.9
Bounds	86.4	87.2 \ 80.1	87.8

Table 3.7: The performance of attribute estimation with the deep features compared to that with HOG feature.

Attribute	HOG	CNN	fc7	fc7 + HOG
Closed Toe	86.30	91.98	90.03	95.60
Back Cover	96.20	96.50	96.60	97.17
Side Cover	92.50	94.24	96.05	96.73
Pointy	92.00	94.55	95.70	96.62
High Thin Heel	89.60	92.90	94.73	95.13
Wedge Heel	94.10	95.53	95.89	98.05
Bounds	86.40	95.78	95.60	96.79

we perform the attribute estimation using our proposed method based on the selected views. Specially, for fair comparison with (Huang et al. [64]), experiments are conducted with view V1, V5 and V6 when $M = 3$. Table 3.6 gives the attribute estimation results with respect to different M . It can be seen that when M increases, with all the six views, the estimation accuracy improves around 3% for most of the attributes. In addition, our propose method performs better than the state-of-the-art when using multi-view shoe images as input ($M = 3$), with over 5% improvement in accuracy for most of the attributes. Also, our proposed method outperforms the state-of-the-art using multi-view shoe images as input. With all the six views, the estimation accuracy improves around 3% for most of the attributes.

Due to the overwhelming discriminative power of the deep feature representations learnt by CNN-based network, we combine it with the HOG feature adopted in the current system. We train an independent CNN model for each of the 7 different types of semantic shoe attributes.

Images in the training and testing set are sampled from different viewpoints. Table 3.7 lists the results of attribute estimation utilizing the concatenated feature vectors and several baselines with deep features compared to that with HOG feature. Here “CNN” indicates we utilize the softmax classifier for the attribute prediction and “fc7” corresponds to the deep features activated from the fc7 layer (the last fully connected layer of AlexNet [75]), which are further fed into SVM classifier for prediction. From the experimental results, it can be seen that the attribute prediction accuracy has a large improvement with the deep feature vectors and the concatenation of the deep feature and the hand-crafted feature could further boost the system performance.

3.4.3 Performance of Shoe Retrieval

In this section, we evaluate the performance of shoe retrieval using the Normalized Discounted Cumulative Gain at top K (NDCG@K) [67], which is a popular evaluation metric in the information retrieval community. Specifically, given one of shoes from the 445 queries, we retrieve the top K similar shoes from the 4000 shoe dataset. The NDCG is computed as follows

$$NDCG@K = \frac{1}{Z} \sum_{i=1}^K \frac{2^{rel(i)} - 1}{\log(1 + i)} \quad (3.6)$$

where Z is the normalization term to ensure the correct ranking score to be one, $rel(i)$ calculates the number of common attribute values between the query index vector and that of shoes in the dataset.

For each of the 445 query shoes, we select $M = 1, 3$ and 6 views as input of the proposed system. While the shoe indexes of the 4000 shoe dataset are computed based on all the six views of the shoes. For $M = 3$, we take the same experimental setting as that mentioned in the attribute estimation. Fig. 3.5 presents the shoe retrieval results in terms of NDCG scores

Fig. 3.5: Performance of shoe retrieval (in terms of NDCG scores) using our proposed system and the state-of-the-art system [64].

for different M , where the performance of the state-of-the-art [64] is also given for $M = 1$. It can be seen that our proposed system achieve higher NDCG scores when compared with the state-of-the-art work for shoe retrieval. Furthermore, it is helpful to have more query views to perform the shoe retrieval. A more intuitive comparison of retrieval results with both the single-view and multi-view shoe images ($M = 3$) as query is demonstrated in Fig. 3.6. It can be seen that our proposed method generally returns visually and semantically matched retrieval results.

To qualitatively evaluate the performance of our proposed shoe indexing and retrieval system, we also conduct a user study among several volunteers by presenting them the top-10 retrieval shoes and requiring them to give a score between 0-7. Two questions related to the appearance and style of the retrieval results are included as shown below:

1) Does the retrieval system correctly return the similar shoes with respect to the matching attributes? Here because we have defined 7 different types of shoes attributes. For each attribute,

Fig. 3.6: Comparison of retrieval results using our proposed system and the state-of-the-art system [64].

if the query and returned shoe item is matched, score as 1; otherwise, score as 0. Thus if two pair of shoes are fully matched, the volunteer should give a score of 7 in total.

2) Does the overall style of the returned shoes match that of the query at first glance? The score is classified into three grades. Score 0-1 indicates that they are not satisfied with the results; Score 2-4 indicates that the results are acceptable but they still need further improvement in terms of the quality; Score 5-7 means that they think the style for half of the returned results are in accordance with that of the query.

It can be seen that Question 1 is objective while Question 2 is somewhat subjective. Specifically, Question 1 implicitly reflects the attribute estimation accuracy. The average score computed over the involved volunteers are shown in Table 3.8. Generally, based on the rated scores, the proposed shoe indexing and retrieval system achieves higher scores than the state-of-the-art [64]. In agreement with the NDCG curve, the system with the multi-view input query presents is superior than that with only a single query image in terms of matched attributes. Also, the overall styles are also improved with more images of multi-view query.

Table 3.8: The subjective evaluation of our proposed retrieval system with a set of M views for each shoe (shown in bold number) and the comparison with [64] (shown in non-bold number).

Question	Single-View ($M=1$)	Multi-Views ($M=3$)	Multi-Views ($M=6$)	State-of-the-art [64]
Question 1	5.2	5.6	5.8	4.8
Question 2	5.1	5.7	5.7	4.4

Fig. 3.7: Examples of shoe image tagging.

3.4.4 Discussions

Our proposed system can be applied to tag the shoe images, which saves the merchants' effort in creating shoe descriptions. As what have mentioned before, most of the online stores display the shoes in a set of multi-view images. With such a set of multi-view images as input, our system can estimate the shoe attributes accurately. Note that, in this case, there is no need to perform view identification. Fig. 3.7 illustrate several examples of the shoes with the corresponding tags (i.e., the estimated attribute) using our system. For illustration convenience and space constraints, only the shoe image from view V_1 is shown, even though the tags for each shoe are predicted from the multi-view images. It can be seen that the tagging correctly describes each of the shoes.

3.5 Summary

In this chapter, we proposed a novel attribute based shoe indexing and retrieval system based on multi-view shoe images. Given a query shoe with multi-view images, we first identified the

views of the images. Then, we proposed to select a set of relevant views for estimating each attribute which is part-aware. During the attribute estimation, we proposed a view-specific part localization model to localize the shoe part accommodating the attribute, from which proper features are extracted and used for the attribute estimation. The index of the query shoe is generated based on the value of the shoe attributes. The final shoe retrieval is performed based on the index distance between the query shoe and those in the dataset. Experiments on the multi-view dataset and daily shoe dataset are carried out to demonstrate the effectiveness of proposed system. The results show that our system achieves good shoe indexing and retrieval performance.

In spite of the good performances achieved in the shoe indexing and retrieval, there are still several limitations in our proposed method and dataset: 1) Since we only consider 7 types of part-aware attributes on shapes, which might not be discriminative enough to differentiate the subtle differences in the large-scale scenario. In the future we will consider to introduce more attributes such as color and texture, etc; 2) The proposed view-specific part localization algorithm might fail at the part location for shoe images in less structured viewpoints. Also, the predicted bounding box for the heeled shoes is not enough to cover the entire heel and it is not that effective on dealing with flat shoes.

Chapter 4

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

In this chapter, we are dealing with the problem of cross-domain shoe image retrieval¹. Given a daily life shoe photo, our goal is to search for exactly the same shoes from the online store images. It is a more challenging problem compared to the attribute prediction and retrieval for online store images as elaborated in Chapter 3, because the query photo from the street domain (daily life scenario) and the reference photo in the online domain (online shop images) have large visual discrepancies due to the scale and viewpoint variation, cluttered background, and self-occlusion. As reviewed in Chapter 2, deep learning based approaches and attribute-based feature representation leveraging Convolutional Neural Network (CNN) have advanced the development of the instance retrieval.

We are motivated to design a cross-domain shoe retrieval framework incorporating both the CNN architecture and the semantic attributes. More specifically, we propose a new CNN model, the Semantic Hierarchy Of attributE Convolutional Neural Network (SHOE-CNN) and a three-level feature representation for discriminative shoe feature expression and efficient retrieval. The tree-structure architecture of SHOE-CNN as well as its newly designed loss function sys-

¹Part of this chapter was presented in [164, 165, 167].

tematically merges semantic attributes of closer visual appearances to prevent shoe images with the obvious visual differences being confused with each other; the features extracted from image, region, and part levels effectively match the shoe images across different domains. We collect a large-scale shoe dataset composed of 14341 street domain and 12652 corresponding online domain images with fine-grained attributes to train our network and evaluate our system. The top-20 retrieval accuracy improves significantly over the solution with the pre-trained CNN features.

4.1 Introduction

The following scenarios happen quite often in our daily life: The shopper may notice her/his favorite shoes worn on others' feet or come across them in a shop window. However, it poses difficulties to describe the target shoes precisely using the text-based search engine. Therefore, it is essential to develop the visual search system for exactly the same shoes from online shops given daily photos, i.e., instance retrieval. This problem is called *cross-domain* shoe retrieval: The query and reference images are from different domains. We name the shoe photos captured in the daily life scenario *street domain*, and the photos from online shop pictures *online domain*. Different from the retrieval for other fashion items like clothes, the cross-domain shoe retrieval has its unique challenges from scale variation, cluttered background, self-occlusion, and viewpoint variation as illustrated in Fig. 4.1.

The cross-domain shoe retrieval can also be considered as a more challenging type of fine-grained image search problem [149]. Given a query image (e.g., bird), the fine-grained image search aims to find the images belonging to the same subcategory object class as the query, while our cross-domain shoe retrieval is to search the items in a finer manner to return the exact shoes rather than the same category or style of shoes as the query.

Fig. 4.1: Example street domain and their counterpart online domain shoe images. The street domain photos may have scale variation, cluttered background, human feet, different poses and viewpoints of shoes, as well as illumination variation. In contrast, pictures from online shops are captured in clean background, ideal illumination conditions as well as fixed viewpoints.

As introduced in Chapter 2, an effective feature descriptor tailored for the object of interest is the key to instance retrieval and general image descriptors (e.g., GIST [102], color features [66], SIFT [95], LBP [4]) have poor performances on objects with large viewpoint variation. Besides viewpoint, the visual discrepancy from street and online domain shoe images caused by self-occlusion, scale variation, and cluttered background could make the feature distance of images depicting the same shoe from different domains even larger than those of different shoes. Existing image-based instance retrieval systems like [71, 109] achieve excellent performances on objects like buildings and logos (mostly planar objects with distinct features), but they cannot be directly applied to shoes partly due to these cross-domain challenges. This motivates us to develop effective feature representations for minimizing the distances of the same shoe in different domains, while maximizing the distances for different shoes.

In fine-grained image search [149] or classification [17, 157], the main challenge lies in how to differentiate the subtle differences (i.e., color of the wing) between subcategories. Examples include identifying different types of birds, species of flowers or breeds of dogs. Due to the descriptive capability of the semantic attribute at capturing the subtle local part differences for fine-grained images, we also incorporate the prediction of the semantic attribute into our retrieval framework. Our goal is to design an appropriate attribute prediction strategy for a

Fig. 4.2: The hierarchical structure of semantic shoe attributes (in black box) and subordinate attributes (SA, in red box) for toes and heels. By grouping “Square Toe”, “Round Toe”, and “Pointy Toe” as “Closed Toe” (SA1) and “Peep Toe” as “Open Toe” (SA2), the semantic attribute prediction will hardly confuse shoes with visually dissimilar toes. The appearances of the attributes are illustrated in (a). The same applies to heels, as shown in (b).

more robust and discriminative shoe feature representation for retrieval. Semantic part-aware shoe attributes, i.e., the attributes of key parts of shoes such as toe and heel, encode more discriminative information. However, correct prediction of semantic attributes for shoes is not an easy task, especially in the cross-domain scenario, because semantically close attributes might have extremely similar visual appearance. To make this problem trackable, our key idea is to consider the hierarchy of semantic shoe attributes and focus on the most discriminative attributes by merging attributes with closer visual appearance as *subordinate attribute*. The

correct prediction of subordinate attribute is always given higher priority to ensure that shoes with obviously different visual appearances are never confused with each other. We show such an example for “toe” and “heel” in the Fig. 4.2. Despite a focus on shoes in this chapter, it is natural to generalize the concept of the subordinate attribute to other objects with multi-class attribute annotations such as clothes.

We then integrate the Semantic Hierarchy Of attribute and subordinate attribute prediction into the deep Convolutional Neural Network (SHOE-CNN) for discriminative shoe feature learning. A new loss function is defined to penalize the misclassification between different subordinate attributes by adding an extra regularization term in addition to the typical softmax-loss term.

An overview of our cross-domain shoe retrieval system is illustrated in Fig. 4.3. Given a query shoe image captured in the daily environment (street domain), we first represent it using features extracted from three levels: the whole image (level 1, orange lines and blocks in Fig. 4.3), the top-3 scored region proposals produced by our proposed region proposal selection approach (level 2, green lines and blocks in Fig. 4.3), and shoe image patches detected by DPM (level 3, blue lines and blocks in Fig. 4.3). Then all three-level features are fed forward to SHOE-CNN module, and we use the last convolutional layer (Conv5) and the first fully-connected layer (FC1) from the SHOE-CNN as a three-level visual representation. The final feature vector concatenates these three parts as a comprehensive representation for shoe images. The reference images in the dataset (online domain) follow the similar procedures except that the top-3 scored region proposals are replaced by the whole image, due to their white background. The distances of feature vectors between the query and reference images are compared by the OASIS algorithm. Finally, reference images with large visual and semantic similarities are returned as the retrieval results.

To further deal with the cross-domain challenges from scale variation and cluttered background, we propose a three-level feature representation for shoes as three different scales of

Fig. 4.3: The overview of our proposed cross-domain exact shoe retrieval system (best viewed in color).

input to SHOE-CNN: the first (coarsest) level is the whole image, the second level is the top-3 candidates from a newly proposed region proposal selection algorithm, and the third (finest) level is the detected shoe parts by Deformable Part Model (DPM) [45]. Features activated from the three-levels of SHOE-CNN are integrated to encode both global structures and local patch details, which capture the appearances of shoe images in a coarse-to-fine manner.

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

The major contributions of our work are threefold:

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

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

4.2 Our Approach

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

4.2.1 SHOE-CNN: Semantic Hierarchy of attribute based Convolutional Neural Network

As illustrated in Fig. 4.2, treating different semantic attributes equally loses the discriminative visual features for shoe parts. We utilize the hierarchical structure of semantic attribute grouping and propose a corresponding CNN with novel loss function (SHOE-CNN) to address this issue.

4.2.1.1 Network Architecture

The network architecture is shown in Fig. 4.4. Inspired by the work of Zhang et al. [172] which designed attribute-sharing deep neural network for the human attribute prediction, we

Fig. 4.4: The architecture of our SHOE-CNN network. It consists of five convolution layers and one fully-connected layers shared for all the attributes. However, FC1 branches out T attribute-specific layers (T is the number of semantic shoe attribute types) thus each attribute owns independent FC2 and FC3 layers.

make appropriate adaption based on the AlexNet [76] by adding a tree-structure set of fully-connected layers after FC1 layer to obtain attribute-sensitive features. Our network contains five convolutional layers and one fully-connected layer with shared parameters for all the attributes, while the last two fully-connected layers are not shared. For each attribute, its last fully-connected layer branches out several units, according to the number of potential subclass attributes values. For example, the part-aware attribute “toe shape” takes 4 values so its specific fully-connected layers branches out 4 units, indicating the probabilities of having the corresponding attribute value. We adopt such a structure because different attribute predictions require different features for prediction; for example, detecting whether the toe shape or the heel shape of a particular shoe requires different features.

4.2.1.2 Loss Function

Our attributes mostly belong to the multi-class attribute type (i.e., take multiple real values), so we use the multi-class softmax loss function. Here we take one attribute a_i as an example and the rest attributes follow the same procedure. Assume the training set contains N images $x_i \in \{1, 2, \dots, N\}$ with $y_i \in \{1, 2, \dots, K\}$ as the corresponding attribute labels, where K is the number of values for the specified attribute a_i . In the standard CNN, the loss function is

minimized in order to maximize the posterior probability for the ground truth class label of the given training sample. For the proposed subordinate attribute, the K values can be further merged into M distinct values. For the toe shape attribute, it originally consists of four attribute values ($K = 4$), corresponding to “Square Toe”, “Round Toe”, “Pointy Toe” and “Peep Toe”. Because the appearances of the first three sub-class attributes obviously have physically-closed toe structure, we consider them to be semantically close and term them as one subordinate attribute (SA1 in Fig. 4.2(a)). Meanwhile, the peep toe has an open toe shape thus we term it as another subordinate attribute (SA2 in Fig. 4.2(a)). Therefore, as to the toe shape attribute, the previous four semantic attribute values ($K = 4$) are further clustered as two subordinate attribute classes ($M = 2$). Our key idea is to penalize the mismatching between subordinate attributes by a larger weight, so that the misclassification between “Round Toe” and “Peep Toe” is imposed with larger penalty than that between “Round Toe” and “Pointy Toe”. Let $h_j^{(i)} (j = \{1, 2, \dots, K\})$ indicate the activation value of the node j from the last fully-connected layer (FC3) in Fig. 4.4, then the probability that the given training sample x_i belongs to class j , denoted as $P_j^{(i)}$, is calculated as:

$$P_j^{(i)} = \frac{\exp(h_j^{(i)})}{\sum_{g=1}^K \exp(h_g^{(i)})}. \quad (4.1)$$

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

$$J_0 = -\frac{1}{N} \left[\sum_{i=1}^N \sum_{j=1}^K \delta(y^{(i)} = j) \log P_j^{(i)} \right], \quad (4.2)$$

where δ is an indicator function. The basic idea behind the loss function is that the prediction error for each class is treated equally. However, in our case, the classification error made within each subordinate attribute cluster is smaller than the errors between clusters. Thus, the penalty between subordinate class should be larger than the misclassification made within subordinate

attribute class. As mentioned above, the sub-class attribute values can be further categorized into subordinate attribute classes by merging the semantically close attribute values into large clusters. Here assume that we have the K attribute values which can be further categorized into M subordinate attribute classes. Let $\{\mathbb{G}_1, \mathbb{G}_2, \dots, \mathbb{G}_M\}$ represent M semantically-close clusters, where M is the number of subordinate attribute classes (clusters). Here $\mathbb{G}_t (t = \{1, 2, \dots, M\})$ is a real-valued set containing the attribute values in the t -th cluster. The proposed loss function based on the property of subordinate attribute is formulated as:

$$J = J_0 + J_1, \quad (4.3)$$

with

$$J_1 = -\frac{1}{N} \left[\sum_{i=1}^N \lambda \left(\sum_{j \in \mathbb{G}_1} \delta(y^{(i)} = j) \log P_{\mathbb{G}_1}^{(i)} + \sum_{j \in \mathbb{G}_2} \delta(y^{(i)} = j) \log P_{\mathbb{G}_2}^{(i)} \dots + \sum_{j \in \mathbb{G}_M} \delta(y^{(i)} = j) \log P_{\mathbb{G}_M}^{(i)} \right) \right]. \quad (4.4)$$

From Eq. 4.4 we can see that the probability for the same cluster is shared. The probability of a given sample x_i belonging to the cluster $\mathbb{G}_t$ is denoted as below:

$$P_{\mathbb{G}_t} = \sum_{j \in \mathbb{G}_t} P_j^{(i)}, \quad (4.5)$$

where λ is a regularization parameter controlling the balance between the original loss term J_0 and the subordinate attribute classification term J_1 . As we can see if we set λ as a small values, then the subordinate classification loss term makes slight contribution; otherwise, the effect of sub-class partition is very weak. We set λ as 2 in our experiments based on the validation data.

The partial derivatives of the new loss term J_1 with respect to the output of the last fully-connected layer ($h_j^{(i)}; j = \{1, 2, \dots, K\}$) is computed and the back-propagation is applied to optimize the parameters for the network. The partial derivatives of J_0 is provided in the literature [138], as demonstrated in Eq. 4.6.

$$\frac{\partial J_0}{\partial h_j^{(i)}} = \frac{1}{N} \left[P_j^{(i)} - \delta(y^{(i)} = j) \right], \quad (4.6)$$

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

$$\frac{\partial J}{\partial h_{j \notin \mathbb{G}_t}^{(i)}} = \frac{1}{N} \left[(\lambda + 1) P_j^{(i)} - \delta(y^{(i)} = j) \right], \quad (4.7)$$

and

$$\begin{aligned} \frac{\partial J}{\partial h_{j \in \mathbb{G}_t}^{(i)}} &= \frac{1}{N} \left[-\lambda \left(\frac{P_j^{(i)}}{P_{\mathbb{G}_t}} - P_j^{(i)} \right) + P_j^{(i)} - \delta(y^{(i)} = j) \right] \\ &= \frac{1}{N} \left[P_j^{(i)} \left(\lambda + 1 - \frac{\lambda}{P_{\mathbb{G}_t}} \right) - \delta(y^{(i)} = j) \right]. \end{aligned} \quad (4.8)$$

Then the generalized partial derivatives of Eq. 4.3 is further formulated as:

$$\begin{aligned} \frac{\partial J}{\partial h_j^{(i)}} &= \frac{1}{N} \left\{ \delta(j \notin \mathbb{G}_t) \left[(\lambda + 1) P_j^{(i)} - \delta(y^{(i)} = j) \right] \right. \\ &\quad \left. + \delta(j \in \mathbb{G}_t) \left[\left(\lambda + 1 - \frac{\lambda}{P_{\mathbb{G}_t}} \right) P_j^{(i)} - \delta(y^{(i)} = j) \right] \right\}. \end{aligned} \quad (4.9)$$

Fig. 4.5: The 1st column pictures indicate the query real-world shoe photos and the rightmost three columns illustrate the top-3 region proposals selected by our proposed method. The middle part shows the initial coarse region proposals generated by the EdgeBox algorithm [179], and the top-3 scored candidates are highlighted in green.

4.2.2 Three-Level Feature Representations for Shoes

We design a three-level feature representation for shoe images to address the challenges of scale variation and background clutter. As illustrated in Fig. 4.3, the features are a combination of SHOE-CNN activated deep features from the whole image (Level 1), the top-3 scored region proposals (Level 2) with large likelihood to contain the shoe, and the fine-grained local part patches detected by DPM (Level 3). Obtaining the first and third level features are straightforward, so we focus on introducing our proposed region proposal selection approach for second level representation.

The goal of generating object region proposals is to produce a set of candidate windows for the subsequent detection process [179]. The initially generated region proposals contain a large number of low-quality candidate boxes, with low Intersection-Over-Union (IoU) scores for the annotated object. IoU is calculated by the intersection of the candidate bounding boxes with the ground truth box divided by the union of them. Thus, it is essential to develop a ranking strategy to rank the list of produced region proposals and determine the location of the object based on the top-ranked candidate boxes. We integrate three criteria for evaluating the quality

of the region proposals: 1) the CNN probability score, 2) the confidence scores produced by EdgeBox, and 3) DPM detection score are employed.

1. *Confidence Score by EdgeBox*: We adopt the Edgebox [179] algorithm to generate the initial region proposals for the subsequent proposal selection procedure. The confidence score returned by EdgeBox is denoted as e .
2. *Probability Score of CNN Model*: We use deep CNN as a binary classifier to differentiate whether an image patch belongs to the foreground or the background. The activation of the softmax layer is extracted to encode the probability of the region proposal being a shoe region, denoted as c .
3. *Confidence Score of DPM*: With the ground truth box provided in the training stage, a deformable part model (DPM) model is learned to detect the location of the shoe. The DPM returns a quality score computing the overlap of its detection bounding box with the region proposals, denoted as d .

After obtaining the three types of scores e , c , and d , we apply rankSVM [20] to learn the weights balancing the contributions of these three scores, and then the object proposals are ranked according to the combined score. The details are as follows:

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

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

$$\forall u_{s_k} > u_{t_k} : f(\mathbf{h}_{s_k}) > f(\mathbf{h}_{t_k}), \quad (4.10)$$

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

$$\frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{(s_k, t_k) \in \mathbb{O}} l(\mathbf{w}^\top \mathbf{h}_{s_k} - \mathbf{w}^\top \mathbf{h}_{t_k}). \quad (4.11)$$

where l is a loss function with the form $l(t) = \max(0, 1 - t)$ and C is the trade-off parameter. The learnt $\mathbf{w}$ indicates the relative importance of the confidence scores from three conventional detection models mentioned above. The final concatenated quality score for each image is denoted as $\mathbf{J} = \mathbf{w}^\top \mathbf{H} = \mathbf{w}^\top [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_P]$, where $\mathbf{J}$ is P -dimensional vector, with each element indicating the quality score for its corresponding region proposal. The top-3 scored region proposals are chosen as the final detection results. Fig. 4.5 illustrates the region proposal selection results.

4.2.3 Deep Features Activated from three-Level SHOE-CNN

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

Given an input image, the image is first resized to a fixed size defined by AlexNet and fed forward to the network. The features are extracted from the last convolutional layer (Conv5)

and we apply the 2-level pyramid mean-pooling [60] ($2 \times 2, 1 \times 1$) on the Conv5 feature map with the size (6×6). Two feature outputs are generated with the size, which are equal to $2 \times 2 \times 256 = 1024$ and $1 \times 1 \times 256 = 256$, respectively. Eventually, we obtain a $1024 + 256 = 1280$ -dimensional Conv5 layer feature representation after spatial pyramid pooling (SPP). The deep feature based representation for an image is obtained by combining the activated features from the FC1 layer and the Conv5 layer, thus generating a feature vector with dimension of $1280 + 4096 = 5376$.

The final feature representation for shoes consists of the visual feature-based representation from three levels, so the resulting feature has a dimension of $5376 + 5376 + 1024 = 11776$. Then PCA is performed on the concatenated three-level feature representation and reduce its dimension to 1024-D. Therefore, each query image is described by a 1024-D feature vector after L2 normalization.

4.2.4 Similarity Metric Learning

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

To evaluate the similarities between images from different domains, we use metric learning to learn the similarity measure by the image triplets. An image triplet typically consists of one reference image and another two photos, with one similar to the reference image in terms of category-level label or instance-level object ID label (also indicated as the relevant class) and the other one shares a different label with the reference (indicated as the irrelevant class). In our case, each shoe image is associated with a unique product ID. Thus we can employ the shoe ID label to identify the relevant and irrelevant class indicated as positive and negative samples.

With the available triplet-based side information, the OASIS algorithm [23] is utilized to evaluate the similarity between pairs of shoes on top of the SHOE-CNN activated deep features. We choose to use OASIS because it is the state-of-the-art approach [49] for learning the semantic similarity, which employs the large margin criterion that enables fast training convergence [23] without degrading the performance.

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

$$S_{\mathbf{W}}(p_i, p_j) = p_i^T \mathbf{W} p_j, \quad (4.12)$$

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

We choose p_i as any sampled street domain images from the image set, and both of the images p_i^+ and p_i^- are obtained from the online shoe shop domain shoe pictures. In an image triplet of shoes, (p_i, p_i^+) is the matched pairs, while (p_i, p_i^-) is the un-matched pairs between the street domain and the online product shoe photo domain. We need to minimize the distance of feature vectors of the same shoe from different domains, so we define a loss function for each of the triplet as

$$l_{\mathbf{W}}^t(p_i, p_i^+, p_i^-) = \max\{0, \gamma - S_{\mathbf{W}}(p_i, p_i^+) + S_{\mathbf{W}}(p_i, p_i^-)\}, \quad (4.13)$$

where γ is the margin constant. The loss function penalizes the positive matched pairs that are far apart and dissimilar unmatched pairs with close distance. Our goal is to minimize a global

loss $L_{\mathbf{W}}$ that accumulates the hinge loss over all the triplets in the training stage:

$$L_{\mathbf{W}} = \sum_{(p_i, p_i^+, p_i^-) \in \mathbb{P}} l_{\mathbf{W}}^t(p_i, p_i^+, p_i^-). \quad (4.14)$$

Passive-Aggressive algorithm [30] provides a solution to minimize the loss over $L_{\mathbf{W}}$ and obtain the optimized $\mathbf{W}$.

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

4.3 Dataset Construction

To the best of our knowledge, the UT Zappos50K [156] is the only publicly available shoe image dataset. It is composed of 50,025 catalog images sampled from *Zappos*², which covers 4 major categories. Their images are faced in the left-side viewpoint with clean background. This dataset cannot be used directly for our problem, because 1) the ground truth attribute labels are provided in the form of relative attribute strength comparison based on the image pairs (e.g., with respect to the relative attribute “comfort”, shoe A is more comfortable than shoe B), 2) they do not provide corresponding images from the street domain, and 3) they only provide 4 types of attributes which are not enough to accurately describe the appearances of shoes. So it is necessary to establish a new dataset to train and evaluate our cross-domain shoe retrieval system.

We collected 14341 daily shoe photos of about 5400 items from the street domain and 12652 shoe pictures from the online domain with fixed pose in the left side 45 degree view and

²<https://www.zappos.com/>

clean background. The images are crawled from two large online shopping websites, *Jingdong*³ and *Amazon*⁴. More specifically, with respect to the online domain images, 4907 images are obtained from *Jingdong* while the remaining are obtained from *Amazon*. For each shoe item from *Jingdong*, we access its online store web page that contains both street domain images showing feet trying on that shoe with natural background and their corresponding online shoe images from multiple views with white background; for images Amazon, they only provide 5-7 shoe images with white background and we only take the one with left side 45 degree view. The complete dataset⁵ and example corresponding shoe images from two different domains are illustrated using the collage and close-up example images in Fig. 4.6.

The shoe photos from *Jingdong* are provided with fine-grained attribute annotations, however, the images crawled from *Amazon* don't have such well-defined attributes. Our attribute annotation is initialized from an automatic extraction of the text descriptions next to the crawled images. We then manually re-touch the attributes using the following three steps: 1) Remove attribute types that are not visible from the image such as materials of the sole and shoe inner material types, etc; 2) Re-annotate crucial attribute labels like heel type, toe shape and color, etc; 3) Merge the attribute values that are semantically close and visually too difficult to differentiate, for instance, "beige" and "ivory teeth" can be considered as one attribute named "white". Finally, we obtain 11 types of semantic attributes with 120 values in total. The specified attribute categories and their corresponding example attribute values are shown in Table. 4.1.

To facilitate the process of training and evaluating our proposed region proposal selection model, a subset of shoe images in our dataset (about 2000 daily shoe photos) are annotated with bounding boxes. In the case of a pair of shoes shown in the image, each single shoe is enclosed

³<https://www.jd.com/>

⁴<https://www.amazon.com/>. In this chapter, the reference images from Amazon are the left 45 degree view images sampled from the dataset introduced in Chapter 3.

⁵Available for download at: <https://sites.google.com/view/crossshoe/home>

Table 4.1: Example attribute values in our dataset. The numbers in parenthesis indicate the total number of values for the corresponding attribute types.

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

by a bounding box. Therefore, to our knowledge extend, our dataset is the first large-scale shoe image dataset providing street domain and online domain matching image pairs with semantic annotations.

4.4 Experimental Results

4.4.1 Data Preparation and Implementation Details

As illustrated in Fig. 4.3, the representation for a query image consists of SHOE-CNN activated features with images/patches of three levels. In the following, we first introduce how we split the data for training and evaluating the SHOE-CNN attribute classification model. Then we mention about the data used for training the corresponding models in generating the three-level images/patches. The first level image is not referred, which is the query image. Finally, the data used in the retrieval stage is introduced.

We train a different SHOE-CNN attribute classification model for each domain respectively

Fig. 4.6: The Image collage from our complete dataset: Left picture demonstrates the images from the street domain and right picture shows the photos from the online domain. Example corresponding images of exact the same pair of shoes from both domains are illustrated in close-up views. Each pair of shoes are represented using three images from the street domain (highlighted in red boxes) and its counterpart online shoe images (highlighted in green boxes). Please zoom-in the electronic version for better details.

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

The training and evaluation of the SHOE-CNN are conducted using the MatConvNet toolbox [138]. The convolutional layers of SHOE-CNN are initialized from the AlexNet model [76]. For the tree-structured layers of SHOE-CNN, the weights are initialized from zero-mean Gaus-

sian distribution. The learning rate is set to 0.001 and divided by 10 after 10 epoches. The batch size is set to 256 for each epoch and the SHOE-CNN for both domains converge after about 20 epoches.

The second level images (top-3 scored region proposals) are obtained based on the confidence scores of three models corresponding to EdgeBox [179], CNN-based shoe detector and DPM detection model [45]. The RankSVM model is employed to learn the weights. The training of each detection model and the rankSVM are introduced below:

1. *EdgeBox*: The EdgeBox algorithm with default parameters [179] are used to generate $P = 100$ initial region proposals and the confidence scores.
2. *CNN-based Shoe Detector*: Here deep CNN is utilized as a binary classifier to differentiate whether an image patch belongs to the foreground shoe region or the background. We sample image patches randomly around the ground truth bounding box regions. For the positive images, the cropping is done such that $\text{IoU} > 0.8$. While the negative samples are cropped with $\text{IoU} < 0.2$. Around 800 images are used in the CNN-based shoe detector training phase. For each image, 50 positive and 80 negative cropped images are generated. Thus, we have about 40000 positive and 64000 negative training examples in total.
3. *DPM Detection Model*: The positive images are ground truth of annotated shoe images and the negatives images are sampled from the initially generated region proposals with $\text{IoU} < 0.2$ with the ground truth of annotated bounding boxes. About 500 positive shoe images and 2000 negative shoe images are sampled from our dataset, with which we train a 5-component shoe model.
4. *RankSVM Training*: We randomly choose 200 shoe images to generate the ordered pairs for weight learning. About $M = 2000$ region proposal pairs are randomly sampled for

each shoe image. To evaluate the performance of region proposal selection, the rest shoe images with annotated bounding box are used to evaluate the region proposal selection performance. C is set to 0.001 based on the validation set.

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

4.4.2 Evaluations of the Proposed System

4.4.2.1 Attribute Prediction using SHOE-CNN

The attribute prediction performance is evaluated based on the mean average precision (MAP). Table. 4.2(a) and Table. 4.2(b) demonstrate the results of each method in the street domain and online domain, respectively. Here Chen et al. [26] utilized the combination of traditional golden features (e.g., SIFT, color in the LAB space, etc) and CRF model for attribute prediction. Ozeki et al. [104] and CNN are deep-based methods without incorporating the hierarchical properties of subordinate attribute in the attribute prediction. On average the attribute prediction using the SHOE-CNN (Ours) is 1.2% higher than the tradition CNN in the street domain and about 1.4% higher in the online domain, which demonstrate the effectiveness of subordinate attributes. We find that the proposed SHOE-CNN with our defined subordinate attributes (denoted as Ours in Table 4.2) outperform other compared methods in most of the attributes, except the attribute like “Suitable Age” and “Types” in the online domain, where the attribute prediction accuracy using the simple CNN slightly outperforms SHOE-CNN. It is mostly due to the data distribution is relatively biased in these two categories of attributes.

For example, for the attribute “Suitable Age”, the number of images having the attribute value “Youth” is about 9 times more than that of the second largest attribute value “Middle Age”.

To further evaluate the effectiveness of the defined subordinate attribute, we vary the number of subordinate attribute M (in Section 4.2.1.2) for each part-aware multi-class attribute as follows: 1) Two subordinate attributes ($M = 2$): This is based on whether the given image has a particular property or not. For example, we just identify that the toe of the shoe is open or closed rather than consider the specific shape of it; and for the “Heel Shape”, we just consider whether the sole is flat or not. In this way, all the part-aware multi-class attributes are clustered into two SAs. 2) Three subordinate attributes ($M = 3$): For the above two subordinate attributes, we further partition the larger SA subset into two SAs. Like the “Toe Shape”, the original SA, composed of “Square Toe”, “Round Toe” and “Pointy Toe”, can be further classified into two SAs. One indicates the non-pointy toe, corresponding to “Square Toe” and “Round Toe”; while the other SA refers to “Pointy Toe”. 3) Four subordinate attributes ($M = 4$): Both the two SAs obtained from 1) are further partitioned, thus the particular part-aware multi-class attribute is classified into four SAs. The results shown in Table 4.2 demonstrate that our definition of subordinate attributes outperform other settings of subordinate attributes. The complete definition of subordinate attribute groupings with different values of M as well as our definition are provided in the supplementary material.

4.4.2.2 Evaluation of the Generality of the SHOE-CNN

We perform experiments using the clothing dataset [26] with semantic attribute labels to evaluate the generality of the proposed SHOE-CNN. It is worth noting that as the SHOE-CNN takes the semantically-close visual attribute into consideration, it can also be generalized other objects with the multi-class attribute annotations. We select three multi-class attribute types from [26] for evaluation, “Neckline”, “Sleeves Length” and “Category”. For the attribute type

Table 4.2: Attribute prediction of the proposed SHOE-CNN compared with other approaches and different settings of subordinate attributes.

(a) Street domain photos							
Attribute Name	Chen et al. [26]	Ozeki et al. [104]	CNN	SHOE-CNN			
				$M = 2$	$M = 3$	$M = 4$	Ours
Close Style	55.10	57.97	60.29	60.10	60.42	60.40	60.58
Toe Shape	57.96	61.72	68.21	69.03	68.62	68.50	69.33
Heel Shape	60.49	60.33	62.38	62.47	63.75	63.49	63.36
Patterns	74.60	76.43	78.02	78.04	78.85	78.16	79.75
Elements	26.92	27.94	27.61	27.63	28.33	28.36	29.11
Styles	30.96	35.73	38.12	35.20	35.74	35.64	35.78
Types	51.06	47.07	52.13	53.18	53.35	53.30	55.35
Suitable Age	82.09	83.54	84.05	84.88	83.85	82.45	85.13
Shoe Width	65.45	66.78	67.62	68.28	68.88	68.57	69.29
Heel Height	52.17	50.19	57.17	57.32	57.51	57.89	58.03
Color	55.08	64.10	67.80	68.43	68.93	68.54	70.17

(b) Online domain Photos							
Attribute Name	Chen et al. [26]	Ozeki et al. [104]	CNN	SHOE-CNN			
				$M = 2$	$M = 3$	$M = 4$	Ours
Close Style	64.64	71.65	73.04	72.95	71.76	72.56	73.90
Toe Shape	71.00	80.48	81.11	82.45	81.63	81.24	84.32
Heel Shape	70.19	73.68	74.32	75.22	77.31	75.17	79.26
Patterns	83.57	83.52	84.91	84.15	84.05	84.86	85.10
Elements	28.24	32.44	34.79	34.28	35.03	34.88	35.26
Styles	32.08	35.77	38.02	38.21	38.06	38.72	38.35
Types	52.77	58.12	62.08	58.38	59.04	59.80	60.42
Suitable Age	90.31	88.70	90.89	89.70	90.22	89.84	90.84
Shoe Width	68.76	71.20	72.62	72.76	73.61	73.37	74.56
Heel Height	62.74	64.57	65.40	66.85	67.32	67.69	67.92
Color	66.72	78.47	78.55	81.67	81.10	81.15	82.06

such as “Neckline”, we consider “Round Neckline” and “V-Shape Neckline” as one SA and the rest as another SA. Table 4.3 demonstrates the comparison results, and we can find that the proposed SHOE-CNN improves consistently over the conventional CNN by 2.3% for each of the three multi-class attributes. Moreover, the attribute prediction accuracy by SHOE-CNN outperforms that using traditional hand-crafted features by 3.3%, 3.4%, and 7.7%.

4.4.2.3 Shoe Region Proposal Selection

Based on the confidence scores from the three models and the learnt weights, a list of ranked region proposals are generated. The detection performance is calculated in terms of the IoU s-

Table 4.3: Comparison of the multi-class attribute prediction accuracy on clothing attribute dataset [26].

	[26] without pose	CNN	SHOE-CNN
Neckline	53.2	54.2	56.50
Sleeves	67.8	69.0	71.20
Category	47.5	53.4	55.20

Table 4.4: Shoe detection performance of real-world shoe photos comparing with two baselines: CNN detection and DPM detection.

Method	IoU Score
DPM [45]	53.22
CNN Detection	58.31
Proposed Method	73.15

core between the selected region proposals and the ground truth bounding box. The top 3 region proposals for each shoe image are kept for the following experiment. We use the DPM detection and CNN detection as the baseline comparison to verify the effectiveness of our proposed region proposal selection scheme. From the results shown in Table. 4.4, it can be seen that our proposed method improves about 20% with respect to the original DPM detection and outperforms the CNN detection by about 15%.

4.4.2.4 Cross-Domain Shoe Retrieval

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

We compare our proposed method (Level 1 + Level 2 + Level 3 feature (w/ SA) + PCA + Similarity Metric Learning) with both the traditional state-of-the-art pipelines and the deep

Table 4.5: Top- K Retrieval Accuracy of our system and baseline results.

Feature Configurations	Dim	Top-1	Top-5	Top-10	Top-20	Top-30
Gist feature [102]	1,024	4.12	7.03	8.31	10.03	11.41
DSIFT + Fisher Vector [69]	1,024	11.11	13.92	17.31	18.54	22.15
Deep Pre-trained CNN (WI)	4,096	11.87	19.22	23.94	28.28	31.59
Metric Network [57]	4,096	21.60	36.65	42.86	48.63	52.42
Siamese Network [12]	4,096	24.02	43.79	52.10	60.40	64.96
Robust Contrastive Loss [142]	4,096	30.88	52.08	59.44	66.69	69.89
Level 1 feature (w/o SA)	5,376	16.14	25.29	29.83	35.53	38.92
Level 1 feature (w/ SA)	5,376	18.21	26.33	31.85	37.02	41.31
Level 2 feature (w/o SA)	5,376	15.37	23.30	27.17	32.92	36.61
Level 2 feature (w/ SA)	5,376	19.34	27.44	33.04	38.76	42.90
Level 3 feature (w/o SA)	1,024	15.23	22.76	26.84	31.68	34.69
Level 3 feature (w/ SA)	1,024	16.04	25.23	31.45	35.17	38.40
Level 1 + Level 2 + Level 3 feature (w/o SA)	11,776	22.37	37.58	43.70	49.49	52.98
Level 1 + Level 2 + Level 3 feature (w/ SA)	11,776	25.19	40.69	47.74	53.77	57.20
Level 1 + Level 2 + Level 3 feature (w/ SA) + PCA	1,024	26.47	44.41	50.67	57.62	60.94
Level 1 + Level 2 + Level 3 feature (w/ SA) + PCA + Similarity Metric Learning	1,024	31.97	49.87	60.29	69.25	73.33

learning based approaches. Two representative instance retrieval approaches with traditional features are used as baselines for comparison: 1) GIST feature [102] with 1024-dimension; 2) Dense SIFT feature followed by fisher vector encoding (DSIFT + Fisher Vector) [69] with the codebook size set as 64. The deep features extracted from the pre-trained CNN network on the whole image are used as a baseline to demonstrate the advantage of our proposed SHOE-CNN over the conventional CNN architecture. We also compare each level feature from SHOE-CNN (w/ SA) with that extracted from fine-tuned CNN without SA (w/o SA). The performance of multi-level deep feature representation is compared to that with the features from different levels individually. To further evaluate the effectiveness of our proposed system, we also compare it with several recent works on the cross-domain product retrieval [12, 57, 142]. Kiapour et al. [57] modeled the similarity relationship between the query and shop feature descriptors through a three-layer fully-connected network. In [12, 142], the siamese deep network is adopted with the contrastive loss and its improved version, robust contrastive loss. Note that all the baselines except [57] utilize the cosine similarity as the metric.

Table 4.5 presents the results of our method and several baselines with different values of

Fig. 4.7: Example retrieval results with top-8 returned shoe candidates. The correctly retrieved results with exactly the same shoe to the query are highlighted in green.

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

top-ranked results.

Finally, we show example retrieval results in Fig. 4.7. The top three rows indicate the successful retrieval results while the last two rows show the error cases. The results demonstrate that our system is capable of handling images with human skin like the second example, some challenging pose (e.g., the first example), also the background clutter (e.g., the newspaper beside the wedge shoes in the third row). From the failure examples, we notice they are caused by large variation of viewpoints (e.g., only the back view of the shoe appear in the fourth row), and negative impact of skin. Also it is still difficult for our proposed feature representation to capture the subtle decoration details (e.g. the number of buckles in the query image in the fifth row). Based on extensive observations of our retrieval results, we find that in most cases we can achieve promising performance on shoes which only occupy a small area of the query image and cluttered background; however, we fail at the query images with large area of human skin exposed.

To provide an intuitive understanding about what the SHOE-CNN learns, Fig. 4.8 illustrates the activated feature map with different levels of query image as the input. It can be seen that the feature maps for the whole image (Level 1) show strong activations on the overall structure of the shoes thus conveying the global information. For the feature maps with the selected high-scored region proposals (Level 2) and local parts (Level 3), generally they have strong activation values on the semantic parts of the shoes that convey the local appearances of the shoes and distinctive features (such as heel in the bottom row) for a unique type of shoes. Such learnt multi-level feature representation benefits the SHOE-CNN in describing both the global and local details of shoes.

Fig. 4.8: SHOE-CNN feature maps. Row 1 and 3: two examples of the daily shoe photos (wedge and high-heel shoes). Row 2 and 4: corresponding activated feature maps from the pool5 layer for the whole query image (Level 1), the corresponding top-scored region proposals (Level 2) referring to each single shoe of a pair, and the local fine-grained patches (Level 3).

Figure 4.9: (a) Detection performance with different number of selected region proposals. (b) Mean average attribute prediction accuracy with respect to different values of λ . (c) Top-20 retrieval accuracy when we vary the parameter γ .

4.4.2.5 Parameter Discussion

To further figure out the influence of parameter settings to the performance of our system, we evaluate the impact of the number of the selected region proposals on the detection accuracy, different settings of λ (Eq. 4.4) on the mean average attribute prediction accuracy, and γ (Eq. 4.13) on the top-20 retrieval accuracy, respectively, as shown in Fig. 9. Note that $\lambda = 0$ indicates the subordinate attribute is not considered. Fig. 4.9(a) demonstrates that the top-3 scored region proposals have the best detection performance compared to other values of the selected region proposals, possibly because the top-3 scored region proposals can tolerate the false detection cases without compromising on the detection accuracy. However, with the number of selected region proposal further increasing, the irrelevant image patches are also included, which may result in accuracy drop. It can be seen from Fig. 4.9(b) that on average incorporating the subordinate attribute improves attribute prediction accuracy. We notice that $\lambda = 2$ achieves the best results. However, as λ further increases, the prediction accuracy begins to drop. It indicates that the common semantic attribute prediction is also helpful to achieve good attribute prediction results and we should balance that with the subordinate attribute prediction. As shown in Fig. 4.9(c), the metric learning significantly boosts the retrieval performance (as shown in Table 4.5) and the best performance is achieved with $\gamma = 0.1$.

4.4.2.6 Computational Complexity

The experiments of the proposed cross-domain shoe retrieval system are performed on a PC with an Intel Xeon E5-2630 CPU with 6 cores, 96G RAM, and a Tesla K40 GPU. For the shoe region proposal selection part, the training for the CNN-based shoe detector takes about 3.5 hours and the DPM model costs about 2.2 hours; for the shoe retrieval part, the training for SHOE-CNN takes about 5.8 hours and 4 hours for the online domain and the street domain, respectively. Given a query image, the proposed region proposal selection algorithm returns

the top-3 scored high-quality region proposals in about 0.73s (Edgebox: 0.15s, DPM: 0.48s, CNN-based shoe detector: 0.1s); given the top-3 scored region proposals, it costs about 0.26s for feature extraction and shoe retrieval in the retrieval step.

4.5 Summary

In this chapter, we present a cross-domain shoe retrieval system which aims at finding the exact same shoe image in online shops given a daily life query image. The SHOE-CNN network with a newly designed loss function by investigating the hierarchical properties of shoe attributes is proposed to extract a three-level feature representation for shoes. We show that our proposed three-level feature representation with the SHOE-CNN activated deep feature outperforms other feature configurations. Moreover, we also build a large-scale shoe image dataset consisting of 14341 street domain and 12652 online domain shoe images. Experiments carried on our dataset demonstrate the effectiveness of our proposed retrieval system. Comparing features extracted from pre-trained AlexNet, the accuracy of our system improves over 40%. There are still several directions for further improvements for which we consider as our future work: 1) To be more consistent with the real-world applications, we will expand our dataset by crawling more images in our reference image set to include shoe photos from different views; 2) The performance of our retrieval system degrades when the human skin is exposed, especially when the feet is partially wrapped by shoes, such as the query shoe image in Fig. 4.3. Thus, we need to develop a human skin detection and exclusion technique to mitigate the negative effect of human skin without losing the details of shoes.

Chapter 5

DeepShoe: A Multi-Task View-Invariant CNN for Model-to-Shop Shoe Retrieval

In this chapter, we are addressing the model-to-shop shoe retrieval¹, which is also referred as the cross-domain shoe retrieval mentioned in Chapter 4. Here the model image shares the same concept with the street domain image indicated in Chapter 4, with the shoes worn on the human model captured in the half-controlled environmental settings. The approach introduced in the previous chapter doesn't explicitly consider the properties of viewpoint variation and local fine-grained differences of shoes into the feature designing process, which might fail at retrieving the same shoes in different viewpoints and return inconsistent results. To address this problem, we not only expand our dataset used in Chapter 4 with shoe images in multiple viewpoints but also incorporate the viewpoint invariance and hard example image pairs which differ in local fine-grained details during the feature learning stage.

As introduced in Chapter 2, the triplet network has gained its popularity in jointly learning the feature representation and similarity metrics in the instance retrieval problem. Built on the effectiveness of the triplet network, a multi-task view-invariant convolutional neural network (MTV-CNN) is proposed to learn the feature embedding for shoes and handle the discrepancy of different scenarios. Specifically, we propose a new loss function that minimizes the distances

¹Part of chapter was presented in [168, 169].

between images of the same shoes captured from different viewpoints. Moreover, we also define a novel definition of shoe style based on combinations of part-aware semantic shoe attributes and develop the associated style identification loss. To train the proposed MTV-CNN in a more efficient way, an attribute-based hard example weighting and mining strategy is developed to smartly select the hard negative examples and anchor images. To evaluate the performance of the proposed MTV-CNN, we collect a large-scale multi-view shoe dataset with semantic attributes (MVShoe) from the daily life and online shopping websites. Experiments on the dataset show our excellence over other compared approaches, which demonstrate the effectiveness of our proposed method.

5.1 Introduction

The challenges in the cross-domain shoe retrieval have been summarized in Chapter 4, which are reemphasized here due to slight differences in the motivation of designing the feature learning network. I) cross-domain differences. It can be seen from top two rows in Fig. 5.1 that even the exact-matched shoe items (image pairs in the dashed box) have large visual discrepancies in the background, viewpoint, scale, degree of occlusion, illumination, etc; II) subtle differences in appearances. The shoes belonging to the same style might differentiate each other with minor fine-grained differences. Thus, the problem of the exact match item retrieval poses more difficulties than searching for the similar item since some items look highly similar that it is also quite difficult for human to tell the differences. For example, the third row in Fig. 5.1, the appearances of the two sets of examples are differentiated by the shoe strap and the heel shape; III) viewpoint variation for the same shoe item. The instance of shoes is a new object that has its own unique challenges as demonstrated in the last row of Fig. 5.1. Other fashion products (e.g., clothing, handbag) are usually shown in the frontal view, however, we often see shoes displayed in a variety of viewpoints in both the online stores and daily scenarios. Although the

system is capable of finding the exact matched online store images with similar viewpoint as that of the query, it may still fail to search for the same item with a less similar view.

Fig. 5.1: Illustrations of main challenges lie in the model-to-shop shoe retrieval. The top two rows (Challenge I) show pairs of exact-match model and online store shoe images with scale, viewpoint, illumination and occlusion changes. The middle row (Challenge II) demonstrates the different shoes with very small fine-grained differences. In the last row (Challenge III), the leftmost column of (c) is the query daily shoe photo, we can retrieve the shoes in the green dashed box with similar viewpoints as that of the query image. However, the system fails at finding the shoe images in the red dashed box in less similar viewpoints.

The aforementioned challenges motivate us to design an effective shoe feature representation that satisfies the following three conditions:

- Domain-invariant: capture the relationship between the exact-matched model and shop scenario photo pairs;
- Semantic and discriminative: incorporate the high-level semantic knowledge and differentiate the fine-grained visual appearances;

- View-invariant: cluster different viewpoint images for the same shoe.

²To learn such semantic-aware view-invariant feature representation for shoes, we propose a novel multi-task view-invariant CNN (MTV-CNN). The distances of features of the same shoes are minimized and those of different shoes are enlarged via the triplet network (domain-invariant, tackling challenge I). To capture the local fine-grained details of the shoes, we incorporate another two auxiliary tasks, attribute prediction and style identification. The “shoe style” is a novel concept by considering different combinations of part-aware shoe attributes (semantic, addressing challenge II). In order to learn a discriminative feature representation that differentiate shoes with similar low-level features but with different high-level semantics, we develop an attribute-based hard example weighting and sampling strategy to smartly select the hard negatives and anchors image in the triplet examples. The triplets are weighted by the hierarchical properties of the multi-class attribute and higher weights are assigned to harder examples (discriminative, handling challenge II). Moreover, we simultaneously pull the feature embeddings of different viewpoint images in the positive bag closer by integrating another auxiliary task of view invariance (view-invariant, solving challenge III).

A multi-view shoe dataset with semantic attribute annotations (MVShoe) is built upon the cross-domain dataset in Chapter 4 with the daily shoes photos and online store images in multiple viewpoints for training and evaluation. Moreover, to increase the diversity in the data, images in more categories such as sport shoes, flat shoes are included. Experiments on this dataset demonstrate the effectiveness of our proposed method for learning the shoe representation. Moreover, it also shows the essences of each module in our proposed network.

Our contributions can be summarized as the following:

1. We develop a multi-task view-invariant CNN that minimizes the distances between images belonging to the same item but from different scenarios (conquering challenge I);

²Note that in this chapter, we are more focusing on developing the view-invariant feature representation.

2. We incorporate the high-level semantic attributes and design an attribute-based hard example weighting and mining strategy to differentiate the subtle differences in appearances (conquering challenge II);
3. The different view images of the same shoes are treated as positive bag and their mutual distances are minimized to pull them closer (conquering challenge III);

The following section is organized as follows. The framework of the proposed MTV-CNN and the details of the corresponding modules are described in Section 5.2. The statistics of the MVShoe dataset are presented in Section 5.3 followed by the implementation details of our techniques in Section 5.4. Finally, the experimental results and discussions are demonstrated in Section 5.5 and the conclusions in the last section.

5.2 Model-to-Shop Shoe Retrieval

In this section, we describe the main components of the proposed model-to-shop shoe retrieval system in more details. The overall framework is first presented, then we introduce the cross-domain shoe retrieval using the conventional triplet network. Based on the triplet loss, we develop the multi-task view-invariant CNN (MTV-CNN) for shoes. Finally, we elaborate the attribute-guided hard example weighting and sampling strategy which incorporates the semantic attribute hierarchy into the design procedure.

5.2.1 The Overall Framework

The overall framework of our model-to-shop shoe retrieval approach is demonstrated in Fig. 5.2. The proposed shoe retrieval method takes the tuplet, including the triplet examples (anchor, positive and negative cropped images) and the positive bag (images of the same shoe

Fig. 5.2: The framework of our model-to-shop shoe retrieval. Here a, p, n indicate the 512-d feature vectors for the anchor, positive and negative image, respectively. The feature representations for the different viewpoint images of the anchor image, named as positive bag, are denoted as p_1, p_2, p_3 .

item as the anchor, in different viewpoints) to train the MTV-CNN. Note that the cropped images through the chapter are the sub-images detected by Fast R-CNN model [50]. The semantic attribute label of the anchor image is also used as the input to compute both the attribute prediction L_a (see Section 5.2.3.1) and style identification loss L_s (see Section 5.2.3.2), so that the learnt feature embedding is supposed to carry semantic information. Moreover, a viewpoint-invariant feature representation is achieved by optimizing the view-variant loss L_v (see Section 5.2.3.3) with images in the positive bag. Thus, multiple tasks are optimized jointly to obtain a semantic view-invariant deep shoe feature representation. During the training procedure, hard examples are weighted based on the hierarchical grouping of the attributes, the loss of which is

denoted as L_t (see Section 5.2.4), and the associated hard negative sampling strategy is further developed for the fine-tuning of MTV-CNN.

In the online retrieval stage, given a real-world query shoe photo, firstly we utilize Fast R-CNN [50] to predict the shoe bounding boxes and the top-2 scored candidate boxes are returned as the region of interests (ROIs). Subsequently, they are fed forward into the MTV-CNN for query feature extraction. For images in the reference gallery, the same procedures are performed. Finally, the Euclidean distance is computed between the query and reference gallery images and is used to return the rank list.

5.2.2 Model-to-Shop Shoe Retrieval Using Triplet Network

Model-to-shop shoe retrieval can be considered as exactly the same shoe item matching problem. We aim at learning an effective embedding and a similarity metric so that images of the same shoe item are clustered and images from different shoes are pushed away. To this end, the triplet-based convolutional neural network is employed for the similarity learning. Conventional triplet-based convolutional neural network consists of three branches with shared parameters and structures, which are fed with two images of the same shoe item and another image from another shoe item.

Let us denote a training set as $X = \{x_i^k | i = 1, 2, \dots, N\}$, where N is the total number of shoe items and x_i^k indicates the k -th image of the i -th shoe item. Given a triplet of shoe images $\{x_i^a, x_i^p, x_j^n\}$, x_i^a (anchor) and x_i^p (positive) have the same shoe item label, while x_j^n (negative) belongs to another item. Here the anchor is randomly selected from the training shoe item set X , the positive image shares the same shoe item ID as the anchor while the negative image belongs to a different shoe item. Fig. 5.3 illustrates the basic architecture of the triplet network. We wish to learn a deep embedding network $f_w(\cdot) \in \mathbb{R}^D$, which maps an input image x to a point in the D -dimensional Euclidean space. Here w is the weights and

biases of the L -layer deep embedding network. The goal of the function $f_{\mathbf{w}}(\cdot)$ is to minimize Euclidean distances $d(f_{\mathbf{w}}(x_i^a), f_{\mathbf{w}}(x_i^p)) = \|f_{\mathbf{w}}(x_i^a) - f_{\mathbf{w}}(x_i^p)\|_2$ of the matched pairs (x_i^a, x_i^p) while maximizing the distances $d(f_{\mathbf{w}}(x_i^a), f_{\mathbf{w}}(x_j^n)) = \|f_{\mathbf{w}}(x_i^a) - f_{\mathbf{w}}(x_j^n)\|_2$ between the non-matched pairs (x_i^a, x_j^n) . The loss proposed in the triplet embedding network can be categorized into two groups, corresponding to the ratio loss [128] and the margin loss [123].

5.2.2.1 Ratio Loss

The ratio loss is worked by normalizing the feature distances to have the unit norm rather than the features. A softmax layer is built on top of the feature distances so that they are within the range of $[0, 1]$.

$$l_+ = \frac{\exp\left(d(f_{\mathbf{w}}(x_i^a), f_{\mathbf{w}}(x_i^p))\right)}{\exp\left(d(f_{\mathbf{w}}(x_i^a), f_{\mathbf{w}}(x_i^p))\right) + \exp\left(d(f_{\mathbf{w}}(x_i^a), f_{\mathbf{w}}(x_j^n))\right)}, \quad (5.1)$$

and

$$l_- = \frac{\exp\left(d(f_{\mathbf{w}}(x_i^a), f_{\mathbf{w}}(x_j^n))\right)}{\exp\left(d(f_{\mathbf{w}}(x_i^a), f_{\mathbf{w}}(x_i^p))\right) + \exp\left(d(f_{\mathbf{w}}(x_i^a), f_{\mathbf{w}}(x_j^n))\right)}. \quad (5.2)$$

Thus the triplet-based ratio loss L to be minimized is defined as:

$$L(x_i^p, x_i^a, x_j^n) = 0.5 \times [(1 - l_-)^2 + l_+^2] = l_+^2. \quad (5.3)$$

5.2.2.2 Margin Loss

The margin loss is used as a convex alternative to the 0-1 ranking loss which is infeasible to compute the gradient descent. It is also used in the siamese network [12] which computes the

Fig. 5.3: The basic architecture of the traditional triplet network with three images (x_i^a, x_i^p, x_j^n) as the input.

loss based on the pair of examples that violates the constraint defined.

$$L(x_i^p, x_i^a, x_j^n) = \max(0, d^2(f_{\mathbf{w}}(x_i^a), f_{\mathbf{w}}(x_i^p)) - d^2(f_{\mathbf{w}}(x_i^a), f_{\mathbf{w}}(x_j^n)) + \sigma), \quad (5.4)$$

where σ is the margin threshold that controls the distance between the matched pairs (anchor vs. positive) and the non-matched pairs (anchor vs. negative) within a triplet. If their relative distance is larger than σ , then the loss L is 0; otherwise, the loss is calculated and the parameters of the network are updated based on the gradient decent.

5.2.3 Multi-Task View-Invariant CNN

Compared with specific class label provided in the recognition task, the limited information within a triplet is less descriptive that might degrade the system performance and slow down convergence speed [173]. To address this issue, we fully make use of the attribute-level annotations of shoes and leverage the multi-task learning to train the triplet network with three auxiliary tasks: 1) attribute prediction; 2) style identification; 3) preserving the viewpoint invariance within same shoes (viewpoint invariant loss). The main task is to the pull images of the same shoe item closer while push images of different shoes far from each other (triplet

loss). The goal of training a multi-task triplet network is to learn a more semantic-aware and viewpoint-invariant embedding that captures both the low-level and high-level similarity.

5.2.3.1 Attribute Prediction

Because the fully-connected layer involves a large number of parameters, for ease of network optimization, only the anchor image is fed into the fully-connected layer to compute the attribute prediction loss. The attributes are annotated based on per shoe item. For example, given N_i images (including daily and online photos) of a particular shoe item x_i , they share the attribute labels $a_i = \{a_i^1, a_i^2, \dots, a_i^m, \dots, a_i^M\}$. Through this work, we deal with the multi-class attributes and M is the number of the multi-class attributes. Let $a_i^m = \{a_i^{m,1}, a_i^{m,2}, \dots, a_i^{m,G_m}\}$ denote a vector indicating the attribute membership, where G_m is the number of attributes in the m -th ($1 \leq m \leq M$) multi-class attribute. Note that for each multi-class attribute, only one attribute label is nonzero, which means $\|a_i^m\|_0 = 1$. As the number of training data with specific attribute labels varies, the weighted softmax loss for the task of m -th attribute prediction L_m is formulated as follows:

$$L_m = - \sum_{g=1}^{G_m} \bar{w}_{m,g} \mathbf{1}(a_i^{m,g}) \log p_i^{m,g}, \quad m = 1, 2, \dots, M, \quad (5.5)$$

where

$$p_i^{m,g} = \frac{\exp(z_i^{m,g})}{\sum_{g=1}^{G_m} \exp(z_i^{m,g})}, \quad (5.6)$$

$\mathbf{1}(\cdot)$ is the indicator function and $p_i^{m,g}$ indicates the probability of assigning x_i to the g -th class in the m -th multi-class attribute and $z_i^{m,g}$ is the output from the softmax layer of g -th node. Take the “Toe Shape” (m -th multi-class attribute) as an example, Eq. (5.6) is to compute the probability of the given shoe item x_i belonging to the “Pointy Toe” class (g -th class). The

weights are inversely proportional to the number of training samples defined as below:

$$w_{m,g} = \frac{1}{N_{m,g}}, \quad \bar{w}_{m,g} = \frac{w_{m,g}}{\sum_{g=1}^{G_m} w_{m,g}} \quad (5.7)$$

where $N_{m,g}$ is the number of training samples with the attribute label g in the m -th multi-class attribute. Note that we only consider $N_{m,g} \geq 50$ to compute the respective weights $w_{m,g}$ and $\bar{w}_{m,g}$. For the minority attributes with $N_{m,g} < 50$, its $\bar{w}_{m,g}$ is directly set to 1 for simplicity. Thus, the attribute prediction loss over all the M multi-class attributes is computed as below:

$$L_a = \frac{1}{M} \sum_{m=1}^M L_m. \quad (5.8)$$

5.2.3.2 Style Identification

For each shoe image, the crawled textual descriptions from the web also include the style of the shoes. Nevertheless, due to subjectivity of human’s fashion perception, they are quite noisy and can not be directly utilized to train the network. Furthermore, it is quite challenging to clean the noisy descriptions by manual annotations, which involves extensive human labor to obtain reliable human judgments of style [73], and could also be person biased. Due to these limitations, we develop a novel style definition for shoes, which can guide the network to learn style-sensitive features.

Our shoe style definition is based on the combination of part-aware semantic shoe attributes. Shoe images with similar part structure and appearances are grouped closer to form a style-specific cluster (style class). By considering both the distribution of attribute labels and their appearances within the multi-class attribute, the “Toe Shape” and “Heel Shape” are chosen to represent the shoe styles. Let a^H and a^T denote their attribute sets, respectively. In this scenario, there are $|a^H| \times |a^T|$ different attribute combinations, where $|\cdot|$ is the cardinality of the attribute sets. Each combination is treated as a shoe style class k_s , where $1 \leq k_s \leq |a^H| \times |a^T|$. Formally,

let $s_i = \{s_i^1, s_i^2, \dots, s_i^{k_s}, \dots, s_i^{|a^H| \times |a^T|}\}$ denote the style label vector for each i -th shoe sample and each entry of the s_i can be denoted as:

$$s_i^{k_s} = \mathbf{1}(a_i^{H, k_H}) \times \mathbf{1}(a_i^{T, k_T}), \quad (5.9)$$

where $1 \leq k_H \leq |a^H|$, $1 \leq k_T \leq |a^T|$. Each shoe item belongs to a style with one non-zero value activated in s_i . After filtering out the styles with less than 10 shoe items in the training data, the clean version of the shoe style list has 22 different style classes. The style classification is similar to the attribute prediction, and it branches out 22 nodes after the pool5 layer. Due to the imbalanced distribution of style class labels, the weighted softmax loss for style classification task is formulated and optimized in the same strategy as that for the attribute prediction. Similar as Eq. (5.5) and Eq. (5.7), the style classification loss L_s is to minimize the softmaxloss between the style label s_i and the predictions.

5.2.3.3 View Invariance

The fundamental principle of the traditional triplet loss denoted in Eq. (5.3) is to enlarge the distance between the feature representation of the anchor image and its negative examples meanwhile shorten that between the negative examples. Nevertheless, there is still problem that it doesn't constrain the distances between the online shop images of the same shoe item, which usually face in multiple viewpoints. As a consequence, the features for the online store shoe images of different viewpoints may be scattered in the embedding space. Even we find the counterpart image as the query with similar viewpoint, however, may fail to retrieve other images of the same item but in less similar viewpoints. To deal with the problem, we design a novel viewpoint invariant loss L_v so that are the shop-scenario images in varied viewpoints are drawn closer.

Assume that the i -th shoe is depicted by $X_{i,s} = \{x_{i,s}^1, \dots, x_{i,s}^{N_s}\}$, which contains N_s model shoe images, and $X_{i,o} = \{x_{i,o}^1, \dots, x_{i,o}^{N_o}\}$ which contains N_o online shoe photos. Here we treat the online shoe photos for the same shoes as a positive bag and make the features of the images in the positive bag similar to each other. This can be formulated to minimize the mutual distance of any image pair $(x_{i,o}^j, x_{i,o}^k)$ in the online shoe photo set $X_{i,o}$, which can be written as:

$$L_v(X_{i,o}) = \frac{1}{2 \times n_d} \sum_{j,k}^{N_o} d^2(f(x_{i,o}^j), f(x_{i,o}^k)), \quad (5.10)$$

where n_d is the number of pairs sampled from the positive bag X^o and $d(\cdot)$ is the Euclidean distance. Indeed, it is desirable to feed all the online store shoe images inside the positive bag into the batch for computation. However, due to the memory bottle neck of deep triplet network, three shop-scenario examples are randomly selected to form the mini-batch. Thus we set $n_d = 3$ in our experiments. The concept of our viewpoint invariant triplet loss is illustrated in Fig. 5.4.

As the input of the proposed MTV-CNN is forward with a $(3 + n_d)$ -tuple, to generate more shop-scenario images of different viewpoints for achieving the view-invariant representation, we also augment the data by in-plane rotation and flipping operation. Note that this augmentation technique is only applied to for those shoe items with $N_o < 4$. We assume the rotation between the model image and online shoe image is a relative shift, thus the rotation is merely performed on the online shoe images. In the training stage, each shoe item should have at least three online store shoe images to guarantee that there are enough images to forward into the MTV-CNN for computing the view-invariant loss L_v .

Therefore, the multi-task loss function J over a mini-batch to be minimized is a weighted combined loss of the above mentioned main task and three auxiliary tasks:

$$J_0 = \frac{1}{T} \sum_{i,j \in X}^T [L(x_i^p, x_i^a, x_j^n) + \alpha_v L_v(X^o) + \beta_a L_a + \gamma_s L_s], \quad (5.11)$$

Fig. 5.4: The illustration for the view-invariant loss. In the right figure, the positive examples of the same shoes (green box) are pulled closer. By minimizing the view-invariant loss as an auxiliary task, the anchor (blue box) and positive bag (green box) are clustered while the negative (red box) are pushed away.

where α_v , β_a , γ_s are the weights to indicate the importance of each auxiliary tasks and T is the number of triplet examples in a mini-batch. The conventional triplet loss in Eq. (5.3) is a special case of our multi-task view-invariant CNN where $\alpha_v = 0$, $\beta_a = 0$ and $\gamma_s = 0$. What is worth to mention is that the conventional triplet loss in Eq. (5.3) assigns the equal importance to each triplet example, and the following Section 5.2.4 will introduce how to impose the weights on the triplet examples to compute the weighted triplet loss L_t (as shown in Fig. 5.2) and mine the hard examples in a progressive stage-wise procedure.

5.2.4 Attribute-based Hard Triplet Weighting Strategy

One of the core step for training the proposed MTV-CNN is the hard triplet sampling. Existing works like [123, 144] mainly sample the hard examples based on the low-level feature distance. To further learn a semantic and discriminative shoe feature representation, we design a novel attribute-based hard triplet weighting and mining scheme. Moreover, we define the difficult levels of the triplet examples in a progressive way and select the hard negatives as well as hard anchors in a stage-wise manner. The whole procedure can be summarized into the following three stages:

Stage 1: Random Selection In the first 10 epoches, the MTV-CNN is trained with the multi-task loss J_0 in Eq. (5.11) and the triplet examples are assigned with equal weight. The anchor as well as the negative examples in the triplets are randomly chosen from the training set X .

Stage 2: Hard Negative Weighting and Sampling After the random selection, the feature representations for the training shoe items are computed during a forward pass through the learnt embedding network $f_w(\cdot)$. For the i -th item, its online store image in the left profile view is viewed as the representation for that item denoted as $\mathbf{z}_i$ and $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_N\} \in \mathbb{R}^{N \times 512}$ indicates the embedding feature matrix for all the shoe items in the training set. We use $\tilde{\mathbf{z}} = [\|\mathbf{z}_1\|_2^2, \|\mathbf{z}_2\|_2^2, \dots, \|\mathbf{z}_N\|_2^2]^T$ to denote the column vector of squared norms of all vectors in $\mathbf{Z}$ and the pairwise Euclidean distance matrix $\mathbf{D}_0$ can be efficiently computed as:

$$\mathbf{D}_0 = \sqrt{(\tilde{\mathbf{z}}\mathbf{1}^T + \mathbf{1}\tilde{\mathbf{z}}^T - 2\mathbf{Z}\mathbf{Z}^T)}, \quad (5.12)$$

where each element $\mathbf{D}_0(i, j)$ is the Euclidean distance between i -th and j -th shoe item. Based on $\mathbf{D}_0$, the neighboring affinity map $\tilde{\mathbf{L}}_0$ are computed with each entry $\tilde{\mathbf{L}}_0(i, j)$ representing the j -th closest neighboring item to the particular i -th item in the embedding space. For example, $\tilde{\mathbf{L}}_0(100, 2)$ is the exact shoe item that is close to the 100-th shoe item in the Euclidean space apart from itself. The top 40% closest neighboring items (for each i -th item) of the affinity map $\tilde{\mathbf{L}}_0$ are chosen as the *Rate-1 hard negative pool* $\mathbb{N}_1$, where the negatives are sampled from.

The rational behind our attribute-based weighting scheme is that if the anchor and negative in the triplet has less resemblance in semantic attributes however are close in the feature space, then we rate this triplet with a higher level of difficulty. In other words, the particular triplet is imposed with higher weight. To this end, we incorporate the hierarchical property of semantic attribute labels for hard triplet weighting. Fig. 5.5 demonstrates the hierarchical tree-structured

Fig. 5.5: Hierarchical grouping of the tree-structured semantic shoe attributes for “Toe Shape”.

grouping of the multi-class “Toe Shape” attribute. It is composed of three levels, *Level-0*, *Level-1* and *Level-2*, which indicate the similarity in a finer manner. For the comparison pair (i, j) , the attribute-based weighting matrix W_a in the semantic space is computed as follows:

$$\mathbf{W}_a(i, j) = \sum_{m=1}^{M_p} d_a^m(a_i^m, a_e^m), \quad (5.13)$$

where the lower index $e = \tilde{\mathbf{L}}_0(i, j)$ indicates the j -th closest neighboring item, M_p is the number of part-aware semantic attributes and d_a^m is the pairwise distance of the m -th attribute in the semantic space, defined as follows:

$$d_a^m(a_i^m, a_e^m) = \begin{cases} 0 & \text{if } [a_i^m]_1 = [a_e^m]_1, [a_i^m]_2 = [a_e^m]_2 \\ w_1 & \text{if } [a_i^m]_1 = [a_e^m]_1, [a_i^m]_2 \neq [a_e^m]_2 \\ w_2 & \text{if } [a_i^m]_1 \neq [a_e^m]_1, [a_i^m]_2 \neq [a_e^m]_2 \end{cases} \quad (5.14)$$

If the i -th and e -th item share the *Level-1* attribute, then $[a_i^m]_1 = [a_e^m]_1$; Otherwise, $[a_i^m]_1 \neq [a_e^m]_1$. For the *Level-2* attribute, it follows the same definition.

For the non part-aware attribute like color, we incorporate a shallow attribute-level without considering the *Level-2* attribute because empirically we find that it is good at capturing the color information than the appearances of the part shape. Thus we impose smaller penalization

to compute the color attribute-based distance $d_a^{\tilde{m}}$, defined as below:

$$d_a^{\tilde{m}}(a_i^{\tilde{m}}, a_e^{\tilde{m}}) = \begin{cases} 0 & \text{if } [a_i^{\tilde{m}}]_2 = [a_e^{\tilde{m}}]_2 \\ w_3 & \text{if } [a_i^{\tilde{m}}]_1 \neq [a_e^{\tilde{m}}]_1 \end{cases} \quad (5.15)$$

Based on the weighting matrix W_a , the attribute-based weighted triplet loss L_t over the *Rate-1 hard negative pool* $\mathbb{N}_1$ is expressed as:

$$L_t(x_i^p, x_i^a, x_j^n) = \mathbf{W}_a(i, j) L(x_i^p, x_i^a, x_j^n), \quad i \in X, j \in \mathbb{N}_1 \quad (5.16)$$

and the multi-task loss J_1 with the negatives sampled from $\mathbb{N}_1$ is denoted as:

$$J_1 = \frac{1}{T} \sum_{i \in X, j \in \mathbb{N}_1}^T [L_t(x_i^p, x_i^a, x_j^n) + \alpha_v L_v(X^o) + \beta_a L_a + \gamma_s L_s], \quad (5.17)$$

Stage 3: Hard Anchor and Negative Selection In the final stage of training we not only select the hard negatives but also actively choose the hard anchors. Similar to the previous stage, the Euclidean distance matrix $\mathbf{D}_1$ and the neighboring affinity map $\tilde{\mathbf{L}}_1$, the top 40% of the closest shoe neighboring items are selected to form *Rate-2 hard negatives pool* $\mathbb{N}_2$. Following Eq. (5.13), the distance matrix $W_a(i, k)$ are re-calculated and updated. The hard anchors are selected based on the attribute inconsistency $\overline{W}_a(i)$ between each i -th shoe item and its neighboring items, which is computed as:

$$\overline{w}_a(i) = \frac{1}{T_n} \sum_{k=1}^{T_n} \mathbf{W}_a(i, k), \quad (5.18)$$

where $T_n = \lfloor 0.4N \rfloor$ is the number of hard negative examples per shoe item.

We define the set of hard anchors $\mathbb{P} = \{i | \overline{w}_a(i) > T_0\}$ as the ensemble of those shoe items with $\overline{w}_a(i)$ larger than the median value over all the training shoe items, where $T_0 = T(\overline{w}_a)$ computes the median value of $\overline{w}_a$.

Then the weighted triplet loss over the set of $\mathbb{N}_2$ and $\mathbb{P}$ is:

$$L_t(x_i^p, x_i^a, x_j^n) = \mathbf{W}_a(i, j)L(x_i^p, x_i^a, x_j^n), \quad i \in \mathbb{P}, j \in \mathbb{N}_2. \quad (5.19)$$

and the multi-task loss J_2 with the negatives sampled from $\mathbb{N}_2$ and anchors from $\mathbb{P}$ is denoted as:

$$J_2 = \frac{1}{T} \sum_{i \in \mathbb{P}, j \in \mathbb{N}_1}^T [L_t(x_i^p, x_i^a, x_j^n) + \alpha_v L_v(X^o) + \beta_a L_a + \gamma_s L_s], \quad (5.20)$$

For back prorogation of the network parameters, the gradients of L_t with respect to the output of the softmax layer (l_+, l_-) is computed as:

$$\frac{\partial L_t}{\partial l_+} = \sqrt{W_a} \frac{\partial L}{\partial l_+}, \quad \frac{\partial L_t}{\partial l_-} = \sqrt{W_a} \frac{\partial L}{\partial l_-}. \quad (5.21)$$

The stochastic gradient descent (SGD) algorithm is utilized to optimize the proposed loss over multiple tasks in a mini-batch. The details of the training strategy is illustrated in Algorithm 1. We stopped the training after about 40 epoches and the training curve is almost converged, from which we obtain the discriminative and semantic shoe feature representation.

5.3 Dataset Construction

The newly collected MVShoe dataset is a larger collection of tagged shoe images in Chapter 4 and besides the left-profile viewpoint for each shoe in the reference gallery (the dataset in Chapter 4), we also provide different viewpoint images for each shoe item. Moreover, we also expand the dataset by crawling more multi-view images from *Amazon*³ or *6pm*⁴. As images from *Amazon* or *6pm* are online clean photos with completely white background, we mainly

³<https://www.amazon.com/>

⁴<https://www.6pm.com/>

Algorithm 1 Training of MTV-CNN**Input:** Training set $X = \{x_i\}_{i=1}^N$, the attribute labels $\{a_i\}_{i=1}^N$ **Output:** The parameters $\mathbf{w}$ of MTV-CNN;**Initialization:** The parameters $\mathbf{w}$ from the CNN model pre-trained on ImageNet;

- 1: **for** epoch = 1 to 10 **do**
- 2: Forward randomly sampled triplets (x_i^a, x_i^p, x_j^n) , positive bag $X_{i,o}$ and attribute labels a_i into the network according to **Stage 1** in Section 5.2.4;
- 3: Compute the multi-task loss J_0 according to Eq. (5.11) and back-prorogate the gradients to update $\mathbf{w}$;
- 4: **end for**
- 5: Compute the $\mathbb{N}_1$ and W_a according to **Stage 2**;
- 6: **while** not converged **do**
- 7: Compute the total loss J_1 based on Eq. (5.17) with hard negatives sampled from $\mathbb{N}_1$ and back-prorogate the gradients according to Eq. (5.21);
- 8: **end while**
- 9: Compute the $\mathbb{N}_2$, W_a and $\mathbb{P}$ according to **Stage 3**;
- 10: **while** not converged **do**
- 11: Compute the weighted and total loss J_2 based on Eq. (5.20) with hard negatives sampled from $\mathbb{N}_2$ and hard anchors from $\mathbb{P}$, then back-prorogate the gradients;
- 12: **end while**
- 13: **return** The parameters $\mathbf{w}$;

use them as the reference gallery photos. Apart from the images, the associated meta-data descriptions of the shoe item and the re-annotation of crucial attributes for cleaner annotation are almost the same as those introduced in Chapter 4. Here only those salient attribute types with balanced images in different classes are remained, as shown in Table 5.1. Eventually, we have about 9800 and 31050 images from the model and online shop scenario in multiple viewpoints with annotated semantic attributes. Some example images in MVShoe are demonstrated in Fig. 5.6.

5.4 Implementation Details

We adopt the high performance VGG 16 Layers model [130] pre-trained on the ImageNet [118] for initialization and the experiments are implemented based on the popular deep learn-

Fig. 5.6: Examples of cropped model and shop scenarios images in our MVShoe dataset.

Table 5.1: Semantic Attribute Groups in MVShoe dataset.

Group	Attributes
Color	Black, Blue, Coffee, Golden, Gray, Green Bronze, Brown, Multi-Color, Orange, Purple, Red Rice White, Silver, White, Yellow, Pink Multi-color, Khaki, Pearl Blue
Toe Shape	Round Toe, Pointy Toe, Square Toe, Peep Toe
Heel Shape	Wedge, Flat, Thick, Block, High-Thin, Wine-Shape Cone-Shape

ing platform, Torch [29]. The training details of the proposed MTV-CNN is illustrated in the Algorithm 1. For optimization, Stochastic Gradient Descent (SGD) [19] algorithm is used for back propagation and the mini-batch size is 8 triplets and the corresponding positive bags (48 images). The learning rate is initialized with 10^{-3} , momentum with 0.9 and weight decay of 10^{-4} . Usually, the image are resized to 256×256 and multiple crops with the size 224×224 are extracted for the network training. However, in the case of the shoes, we find that the cropped images lose discriminative details of the part patch appearances, which have a large impact on the overall appearances of shoes. To address this issue, we keep the aspect ratio of the original

image (w, h) and rescale the image such that length of its longer side is 224. Then the scaled shorter side is padded by $224 \times [1 - \frac{\min(w,h)}{\max(w,h)}]$ white pixels. We set $\alpha_v = 0.05$, $\beta_a = 0.05$ and $\gamma_s = 0.05$ in our experiments. For the training of Fast R-CNN, we follow the experimental settings in [50] and the positive examples are those cropped images which satisfy $\text{IoU}^5 > 0.7$; otherwise, they are used as negative examples. The attribute-based weighting parameters in Section 5.2.4 are experimentally set as $w_1 = 0.7$, $w_2 = 1.5$, $w_3 = 0.4$.

5.5 Experiment Settings

5.5.1 Training and Testing Data

Our proposed framework consists of two core components: shoe detection and cross-scenario shoe retrieval based on the detected bounding boxes. For the training of shoe detection module, 2292 images are used to learn the Fast R-CNN model, including the real-world and online shop photos. Each image (usually with a pair of shoes) is annotated with two ground truth bounding boxes indicating each single shoe of a pair. We follow the same parameter setting as in [179] for the input region proposals generated by Edgebox.

For the shoe retrieval procedure, about 3130 shoe items with several daily shoe photos and online store images of different viewpoints (i.e., 24000 images after augmentation) are fed into the proposed MTV-CNN for training. Each shoe item has an associated product item ID that can help us to find matched pairs and non-matches for the network training. About 30200 online store shoe images with multiple viewpoints are used as reference gallery and 4401 daily shoe photos are used as the queries where each of them has a counterpart in the reference set.

⁵IoU is defined as the intersection of the candidate window with the ground truth box divided by the union of them.

5.5.2 Comparison Methods

To analyze the performance of our proposed MTV-CNN, we compare with a variety of approaches. These approaches can be divided into two groups. One group extracts the hand-crafted feature for retrieval, including (1) GIST feature [102] with 512-dimension, (2) Dense SIFT feature followed by fisher vector encoding [69] with the codebook size set as 64, denoted as DSIFT + Fisher Vector, and (3) Hessian affine detector with SIFT descriptor followed by the VLAD encoding [9] with the codebook size set as 128, denoted as HessianSIFT + VLAD. The other group exploits the deep learning methods for feature learning and the learnt deep shoe feature is used for retrieval. This group includes the following competitors: (1) Pre-trained fc6: deep feature activated from the first fully-connected layer of the VGG 16 layers model [130] with the cropped images as input, (2) Pre-trained conv5: deep feature activated from last convolutional layer the VGG 16 layers mode [130] followed by the max-pooling operation, (3) Siamese Network [12]: the hinge loss with the siamese network (4) Triplet Network [128]: the triplet network with the ratio loss in Eq. (5.3), (5) Multi-scale Viewpoint Invariant Triplet Network [169] (MsVI-CNN): the triplet network considers the viewpoint invariance with the WI-CNN trained using the whole images and RP-CNN using the high-quality region proposals, (6) TriNet [61]: a variant of the classical triplet loss, named *batch hard loss*, which is trained with online mining of moderate hard examples within the mini-batch. Note that here the squared Euclidean distance and set margin are utilized for fair comparison.

5.5.3 Evaluation Protocol

To evaluate the performance of our proposed model-to-shop retrieval system, we adopt two evaluation metrics: Top- K Retrieval Accuracy and Normalized Discount Cumulative Gain (NDCG).

5.5.3.1 Top-K Retrieval Accuracy

It computes the percentage of the query images whose top- K retrieval results depicts the same shoe item as the query photos. If the top- K returned results contain the true-match as the query, then it is considered as a successful result; otherwise, it is an error case.

5.5.3.2 Normalized Discount Cumulative Gain (NDCG)

NDCG score is used to evaluate the retrieval results in terms of attribute overlapping between the query image and reference gallery, defined as $\text{NDCG@K} = \frac{1}{Z} \sum_{k=1}^K \frac{2^{r(k)} - 1}{\log(1+k)}$, where Z is a normalization term and $r(k)$ assesses the relevance between the k -th ranked reference gallery image and the query image by computing the number of their overlapping attributes in different levels. Note that because some multi-class attribute (e.g., color) are too fine-grained to differentiate the specific classes so that we consider the attribute-level matching in the Level-1 sense. That is to say, if the query image and the retrieved image share the Level-1 similarity as shown in Fig. 5.5, then they are supposed to be matched for the particular attribute.

5.5.4 Results

In this section, we extensively conduct the experiments to compare the proposed MTV-CNN with the ratio loss as well as the margin loss to other methods. Then we demonstrate the importance of each component within the framework in details, including the contributions of the auxiliary tasks, hard example triplet weighting and sampling. Through the experimental analysis, the results can clearly reflect the impact of each module to the overall system.

Table 5.2: Top-20 retrieval accuracy on the MVShoe dataset.

Method	Top-20(%)
GIST [102]	5.02
DSIFT + Fisher Vector [69]	16.32
HessianSIFT + VLAD [9]	18.77
Pre-trained fc6 [130]	26.52
Pre-trained conv5 [130]	32.27
Siamese Network [12]	58.53
Triplet Network [128]	65.69
Multi-scale Viewpoint Invariant Triplet Network [169]	67.67
TriNet [61]	77.46
MTV-CNN (ratio loss)	75.25
MTV-CNN (margin loss)	79.32

5.5.4.1 Comparisons with Other Competitors

Table 5.2 presents the top-20 ($K = 20$) retrieval accuracy of the proposed MTV-CNN (ratio loss & margin loss) and other compared methods. Our proposed MTV-CNN achieves the best performances among all the competitors, which verify the effectiveness of our framework learning the semantic viewpoint invariant deep shoe feature representation in a multi-task manner. Here for convenience, Top-20 is short for Top-20 Accuracy. From the experimental results we find that the deep features (fc6 and conv5) with the pre-trained model outperform the state-of-the-art hand-crafted feature (HessianSIFT + VLAD) by about 7.8% and 13.5%, respectively. This verifies that the generic deep feature pre-trained on the ImageNet is transferable to the target shoe dataset and also an effective representation for the cross-scenario shoe retrieval task. Furthermore, it can be seen that the retrieval results using the feature activated from the conv5 layer has an improvement of about 5.5% over that extracted from the fc6 layer, which inspires us to utilize the feature extracted from the conv5 layer as the building block for learning the feature embedding network.

We also compare the MTV-CNN with several state-of-the-art works that utilizes the deep

Table 5.3: NDCG score on the MVShoe dataset.

Method	NDCG@20
GIST [102]	0.750
DSIFT + Fisher Vector [69]	0.779
DSURF + VLAD [9]	0.782
Pre-trained fc6 [130]	0.791
Pre-trained conv5 [130]	0.805
Siamese Network [12]	0.890
Triplet Network [128]	0.892
Multi-scale Viewpoint Invariant Triplet Network [169]	0.896
TriNet [61]	0.926
MTV-CNN (ratio loss)	0.942
MTV-CNN (margin loss)	0.938

metric learning, Siamese network [142], the triplet network with margin loss [123], the triplet network with ratio loss [128], TriNet with batch hard loss [61]. We can find the retrieval performance has been significantly boosted with the learnt feature representations fine-tuned on the target shoe dataset. Compared to the two-branch siamese network, the three-branch triplet network achieves better performance by about 6.9% (58.53% vs. 65.69%) and 11.9% (58.53% vs. 70.46%) improvements. One possible reason is that the imposing the constraint between the similar and dissimilar image pairs within a triplet of three examples is more effective at capturing the semantic high-level features of shoes. We also find that the triplet network with the margin loss converges faster than that with the ratio loss and achieves better performance (70.46% vs. 65.69%).

With the further extensions of the proposed MTV-CNN has about 7.6% (ratio loss) and 11.7% (margin loss) improvements over the preliminary version in [169], which demonstrates the effectiveness of the added components in this work (e.g., auxiliary task, weighted triplet loss, hard example mining). We also investigate the performance of our proposed MTV-CNN with both the ratio loss and margin loss. The experiments indicate that the proposed network with the margin loss improves over that with ratio loss by about 4.1%. It is possibly because

that the hinge loss is capable of making use the similar/dissimilar examples within a triplet that violates the constraint and updating the network parameters. Moreover, our approach still achieves better performance over TriNet [61], which demonstrates the effective mining of the hard examples based on the semantic attribute consistency.

As the proposed MTV-CNN utilizes the auxiliary task of attribute prediction and style identification to regularize the network, the learnt deep shoe feature representation is expected to carry semantic information. The NDCG score is used to assess the attribute-level consistency between the query image and the top- K ranked list. Note that the score is in the range of $[0, 1]$ and the higher the NDCG score, the better is the retrieval performance in the semantic sense. The results are reported in Table 5.3 when $K = 20$ and we can see that our method is also the best among the compared methods in terms of attribute-level matching, which indicates our proposed network is able to learn the semantic-aware features. Moreover, the NDCG@20 score is proportional to that of the top-20 retrieval accuracy except for the case of the proposed MTV-CNN with the ratio loss and the margin loss. We can find that the former achieves inferior performance than the latter in the top-20 retrieval accuracy, nevertheless, it is better than the latter in terms of NDCG@20 metric. It demonstrates that the proposed MTV-CNN with the ratio loss generally better at learning more semantic feature representations.

5.5.4.2 Evaluation of Different Returned Items

To give a thorough analysis of the advantage of our method over the compared methods, Fig. 5.7 illustrates the top- K retrieval results with K varying from $[1, 10, 20, \dots, 80]$. We can clearly see that our approach always achieves the best performance over other alternatives with different settings of K . Moreover, the retrieval performance with fine-tuned deep feature through the feature embedding network on the target shoe dataset has a significant improvement compared to the generic deep feature and hand-crafted features. When K reaches to 80, the system is capable of returning the exact-match for almost 90% of the daily shoe photos.

Fig. 5.7: Top@K retrieval accuracy with K varying from 1 to 80.

Fig. 5.8: Example retrieval results shown with the green bounding box indicates the exact-match to the query item.

5.5.4.3 Qualitative Evaluation of the System

To give a more intuitive illustration of the performance of our model-to-shop shoe retrieval system, Fig. 5.8 presents two example daily life query images and the corresponding top-10 retrieved shoe images. For each example, the results in the first row are generated utilizing the MTV-CNN with the ratio loss and the second row are achieved utilizing MTV-CNN with the margin loss. It can be seen that most of returned results are quite similar as the exemplar query images in the global aspects (e.g., style, color) and local aspects (e.g., decorations, strap, buckle). For the first example, the results for two types of losses exactly the same shoes as the query in three viewpoints, which validates the effectiveness of the learnt feature representation at capturing the viewpoint invariance. For the second example, even though the pose of the shoes is quite strange, our system is also able to return visually similar shoes as the query and the ratio loss successfully return the exact match which demonstrates the superior power of the learnt semantic and viewpoint invariant feature representations.

5.5.4.4 Evaluation of Different Deep Architectures

Because the proposed MTV-CNN is not limited to the VGG 16 Layers CNN, other architectures of the convolutional neural network can also be fitted into our approach (e.g., VGG CNN-S). Here we also conduct the experiments of utilizing CNN-S as shown in Table 5.4 for the ratio loss and Table 5.5 for the margin loss in order to analyze the choice of the CNN model to the retrieval performance. The VGG 16 Layers achieves much better results compared to CNN-S model. This also in accordance with other instance retrieval pipeline that achieved the best performance using VGG 16 Layers network, which contain more deeper layers. Again, the experimental results with different CNN structures indicate that the ratio loss exhibits more capabilities of learning semantic attributes compared to the margin loss.

Table 5.4: Comparisons of different CNN architectures on the retrieval performance in terms of Top@20 Accuracy when utilizing the ratio loss.

CNN Models	Top-20(%) (Ratio Loss)	NDCG@20 (Ratio Loss)
VGG CNN-S [22]	49.35	0.893
VGG 16 Layers CNN [130]	75.25	0.942

Table 5.5: Comparisons of different CNN architectures on the retrieval performance in terms of Top@20 Accuracy when utilizing the margin loss.

CNN Models	Top-20(%) (Margin Loss)	NDCG@20 (Margin Loss)
VGG CNN-S [22]	54.12	0.887
VGG 16 Layers CNN [130]	79.32	0.938

5.5.4.5 Impacts of the Auxiliary Tasks

Our proposed MTV-CNN is in the framework of multi-task learning. The network consists of one main task of the weighted triplet loss and other three auxiliary tasks: (1) attribute prediction, (2) style identification, and (3) viewpoint invariance. The multi-task framework benefits from other related auxiliary tasks, in the cross-domain shoe retrieval scenario, we aim to learn a semantic and viewpoint-invariant deep shoe feature representation.

To evaluate the specific contribution of each auxiliary task, the experiments are designed by removing one auxiliary task and evaluate its impact on the overall retrieval accuracy as shown in Table 5.6 (ratio loss) and Table 5.7 (margin loss). We denote the triplet network alone as (TN), the task of attribute prediction as (AP), style identification as (SS) and the viewpoint invariance as (VI). The individual tasks are stacked to the framework one by one and the corresponding retrieval performance is reported. The simplest version is using the triplet network directly. We can find that the retrieval performance improves in terms of top-20 retrieval accuracy gradually as more auxiliary tasks are added to the framework. The baseline module is TN, and the results show that the performance with the combination of the baseline module and the auxiliary task modules TN + AP, TN + VI and TN + SS improve over that achieved by merely the baseline

module (TN), which verifies that each of the task benefits the overall system and the robust shoe feature representation learning. The model using TN + AP + SS + VI obtains 1.6% (72.01% vs. 70.48%, ratio loss) and 1.0% (75.64% vs. 74.60%, margin loss) improvement compared to that without VI module, which indicates the effectiveness of the viewpoint invariant loss.

From the NDCG@20 score metric, we can clearly find that the stacked model TN + AP + SS + VI achieved the best performance and the learnt feature carries the most semantic meaning. Moreover, the attribute prediction TN + AP and the style identification modules TN + SS are helpful for learning semantic feature representation compared to the viewpoint invariant module TN + VI, which is also according to our expectations. By stacking the AP and SS module together, the retrieval performance could further improve, which demonstrates the importance for each components in our proposed MTV-CNN framework.

Table 5.6: Contributions of the auxiliary tasks and each stage to the proposed MTV-CNN for the ratio loss.

Auxiliary Tasks	Top-20(%) (Ratio Loss)	NDCG@20 (Ratio Loss)
TN	65.69	0.892
TN + AP	67.96	0.921
TN + VI	69.71	0.910
TN + SS	68.69	0.919
TN + AP + SS	70.48	0.925
TN + AP + SS + VI	72.01	0.930
Stage 1	67.05	0.930
Stage 1 + Stage 2	73.69	0.938
Stage 1 + Stage 2 + Stage 3	75.25	0.942

5.5.4.6 Impacts of Hard Examples Weighting and Sampling

The hard triplet weighting and sampling strategy are based on the attribute and performed in a progressive way, which can be divided into three stages (e.g., Stage 1, Stage 2 and Stage 3). Here Stage 2 mainly searches for the hard triplets that the distance between the anchor

Table 5.7: Contributions of the auxiliary tasks and each stage to the proposed MTV-CNN for the margin loss.

Auxiliary Tasks	Top-20(%) (Margin Loss)	NDCG@20 (Margin Loss)
TN	70.46	0.906
TN + AP	73.48	0.922
TN + VI	71.67	0.913
TN + SS	72.64	0.921
TN + AP + SS	74.60	0.930
TN + AP + SS + VI	75.64	0.933
Stage 1	70.92	0.925
Stage 1 + Stage 2	76.10	0.932
Stage 1 + Stage 2 + Stage 3	79.32	0.938

image and negative example is small, while Stage 3 looks for hard anchor images which is equivalent to mining for the hard positive images. We systematically design the experiments to demonstrate the importance of each stage. The lower part in Table 5.6 (ratio loss) and Table 5.7 (margin loss) show the retrieval performance of adaptively more stages. We find that the Stage 2 and the Stage 3 improves the retrieval performance in terms of the Top@20 Accuracy by 6.6% and 1.6% (ratio loss) versus 5.2% and 3.2% (margin loss) respectively. It indicates that the importance of both hard negative example (Stage 2) and hard anchor mining (Stage 3). The triplets with hard examples are more helpful for the gradient update and more effective for learning the parameters for the embedding network. Moreover, we can find that both hard negative example mining and hard anchor/positive example selection have increased NDCG score, which means that the selection of the negative and anchor images is of vital importance to the ultimate goal of learning a semantic feature representation.

5.6 Summary

In this chapter, an extended and novel triplet network, the multi-task viewpoint invariant triplet network (MTV-CNN) is presented for learning the shoe feature embedding. We show that

the network learnt in a multi-task manner is simultaneously solving the attribute prediction, style identification and viewpoint invariant loss, and the main task of triplet loss is helpful for learning a semantically-aware deep shoe feature representation. During the training procedure, the hard negative examples are progressively mined and the triplet examples are assigned with different weights based on inconsistency between the high-level semantic attributes and low-level feature distances to penalize more the hard negative examples. The experiments performed on our multi-viewpoint cross-scenario large-scale shoe dataset (MVShoe) show the effectiveness of our proposed method. In the future, we plan to explore more different categories of fashion products and consider the cross-category fashion product matching task.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

Visual analysis of fashion is an important research direction in the computer vision and multimedia community. Shoes, as the necessary item of a person's belongings, play an essential role in person's lives and its online shopping revenues keep increasing. The vision analysis of shoes is still at the beginning stage. Imagine that when a shopper is attracted to a pair of shoes presented in the online store, and she would like to browse similar style of shoes but in more brighter color or with different decorations, higher heel shape, etc. In this case, the in-store shoe retrieval is quite helpful for the customer, which is capable of offering the customers with more choices of shoes favoring to their tastes. Also, the online shoppers also benefit from this kind of framework because it allows them to automatically annotate the properties of the uploaded images, which saves much human labor for manual annotation with millions of new products released every year. In this way, the proposed shoe attribute prediction and in-store retrieval system benefit in both the customers' and the online shoppers' aspects. For sure, they will greatly enrich the consumers' online shopping experience. In this thesis, we study the problem of shoe attribute prediction and the retrieval in the online domain as well as the cross-domain. We investigate different approaches to deal with it: the traditional hand-crafted features as well

as the deep-learning based architecture guided by the semantic attribute learning.

In the first task, we develop a novel attribute based shoe indexing and retrieval system based on multi-view shoe images. Given a query shoe with multi-view images, we first identify the views of the images. Then, we propose to select a set of relevant views for estimating each attribute which is part-aware. During the attribute estimation, we propose a view-specific part localization model to localize the shoe part accommodating the attribute, from which proper features are extracted and used for the attribute estimation. The index of the query shoe is generated based on the value of the shoe attributes. The final shoe retrieval is performed based on the index distance between the query shoe and those in the dataset. Experiments on the multi-view dataset and daily shoe dataset are carried out to demonstrate the effectiveness of proposed system. The results show that our system achieves good shoe indexing and retrieval performance.

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

In the third task, we address the model-to-shop shoe retrieval (i.e., the cross-domain shoe retrieval) via a multi-task viewpoint invariant triplet network, which embeds the feature of images for the same shoes similar to each other. We show that the network learnt in a multi-task manner is simultaneously solving the attribute prediction, style identification and viewpoint invariant loss, and the main task of triplet loss is helpful for learning a semantic-aware deep shoe

feature representation. Our training strategy is a stage-by-stage process, progressively selecting the increasingly challenging examples and assigning the weights to the triplets based on the semantic attributes to penalize more those hard examples. We also establish a novel multi-viewpoint cross-scenario large-scale shoe dataset, namely MVSS. The experiments performed on our newly built dataset achieving about 80% in terms of the top-20 mean average accuracy utilizing our proposed method, far outperforming other compared methods.

6.2 Future Work

Avenues for further study arising from the present research are briefly summarized below. There are still several directions for further improvements for which we consider as our future work:

1. The performance of our retrieval system degrades when the human skin is exposed, especially when the feet is partially wrapped by shoes, such as the query shoe image in Fig. 4.3. Thus, we need to develop a human skin detection and exclusion technique to mitigate the negative effect of human skin without losing the details of shoes.
2. The detected bounding box of the shoes still incorporates complicated backgrounds that are similar to the visual patterns of the shoes (e.g., stripe, lattice), which might degrade the retrieval performance. In this case, more refined region by performing semantic parsing of the shoe image is expected to be used as the complimentary information for shoe retrieval. Thus, we can refer to the area of semantic parsing to further eliminate those un-desired background.
3. The triplet network involved in Chapter 5 is based on the fully supervised attribute labels for accurate prediction as an auxiliary task for the MTV-CNN training. However, the fully supervised attribute labels are not always available or accurate. We aim to investigate how to

make use of the large number of images without attribute annotation and learn the proposed network in a semi-supervised manner. Moreover, we intend to incorporate more attributes so that the styles can be in more choices rather than a combination of three multi-class shoe attributes, which is for sure to further improve the system's performance.

4. One of the unique challenges for the visual analysis of shoes is the large viewpoint variation. Currently, the dataset of the street-domain photos doesn't have very strange viewpoints. We will integrate the 3D model of shoes to identify the shoe images with similar styles and predict its structure information such as orientation, scale, etc. Then based on the inferred structure information, we extract the geometrical/spatial features on the 3D model and appearance features on the 2D image. The spatial information rendered from the 3D model is to help to further boost the performance of both the cross-domain shoe retrieval and in-store shoe retrieval.

5. Currently, the proposed retrieval system can't satisfy the tastes for all the customers. To address this issue, we would like to build a personalized shoe recommendation system by allowing customers to specify which part of the shoes they are satisfied with and which part needs to be further adjusted according to their envisioned shoes. With such part-aware feedbacks in combination with the attributes, the system is capable of presenting the results tailored to the customers' requests.

6. Inspired by fact that the performance of system is to be improved by larger number of images, we could expand the dataset with more synthetic shoe images with subtle fine-grained differences, which are generated from the emerging Generative Adversarial Networks (GANs) [55]. With such specially generated data, our system is able to better distinguish the image pairs that are close in the semantic space.

7. The proposed cross-domain shoe retrieval system still has its limitation when dealing with the shoe images from the wild, with human body in the street with un-constrained poses. To deal with this scenario, the recently proposed DSSD algorithm is adopted for shoe detection in

the wild, which achieves good performances on small objects. Then these small resolution shoe images are fed into the network for training together with the high resolution images crawled from *Jingdong*. To improve the feature representation capability, we will replace the previously utilized AlexNet or VGG-16 with more wider and deeper networks such as GoogleNet or ResNet.

Author's Publications

Journal Papers

1. Huijing Zhan, Boxin Shi, and Alex C. Kot, "Cross-Domain Shoe Retrieval with a Semantic Hierarchy of Attribute Classification Network", *IEEE Transactions on Image Processing*, vol. PP, no. 99, pp. 1-1, 2017.
2. Huijing Zhan, Boxin Shi, and Alex C. Kot, "DeepShoe: An Improved Multi-Task View-Invariant CNN for Street-to-Shop Shoe Retrieval", submitted to *IEEE Transactions On Multimedia*.

Conference Papers

1. **H. Zhan**, B. Shi and Alex C. Kot, "Street-to-Shop Shoe Retrieval", in *British Machine Vision Conference (BMVC)*, 2017.
2. **H. Zhan**, B. Shi and Alex C. Kot, "Street-To-Shop Shoe Retrieval with Multi-scale Viewpoint Invariant Triplet Network", in *proceedings of IEEE International Conference on Image Processing (ICIP)*, 2017.
3. **H. Zhan**, B. Shi and Alex C. Kot, "Fashion Analysis with a Subordinate Attribute Classification Network", in *proceedings of IEEE International Conference on Multimedia and Expo (ICME) (oral presentation)*, 2017.
4. **H. Zhan**, B. Shi and Alex C. Kot, "Cross-Domain Shoe Retrieval using a Three-Level Deep Feature Representation", in *proceedings of IEEE International Symposium on Circuits and Systems (ISCAS)*, 2017.
5. **H. Zhan**, S. Li, and Alex C. Kot, "Tagging the Shoe Images by Semantic Attributes", in *proceedings of IEEE International Conference on Digital Signal Processing (DSP)*, 2015.

Bibliography

- [1] China is top of the online fashion world. <https://www.statista.com/chart/6045/china-is-top-of-the-online-fashion-world/>.
- [2] Online sales of shoes to run far ahead of past brick-and-mortar business. <http://www.supplychainbrain.com/content/industry-verticals/apparel/single-article-page/article/>.
- [3] Online-shopping and e-commerce worldwide: Statistics and facts. <https://www.statista.com/topics/871/online-shopping/>.
- [4] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006.
- [5] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 819–826, 2013.
- [6] Ziad Al-Halah and Rainer Stiefelhagen. How to transfer? zero-shot object recognition via hierarchical transfer of semantic attributes. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 837–843, 2015.

- [7] Anelia Angelova and Shenghuo Zhu. Efficient object detection and segmentation for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 811–818, 2013.
- [8] Relja Arandjelovic and Andrew Zisserman. Multiple queries for large scale specific object retrieval. In *Proceedings of the British Machine Vision Conference*, pages 1–11, 2012.
- [9] Relja Arandjelovic and Andrew Zisserman. All about vlad. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1578–1585, 2013.
- [10] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. Neural codes for image retrieval. In *Proceedings of the European Conference on Computer Vision*, pages 584–599. 2014.
- [11] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. *Proceedings of the European Conference on Computer Vision*, pages 404–417, 2006.
- [12] Sean Bell and Kavita Bala. Learning visual similarity for product design with convolutional neural networks. *ACM Transactions on Graphics*, 34(4):98, 2015.
- [13] Asa Ben-Hur and Jason Weston. A users guide to support vector machines. In *Data mining techniques for the life sciences*, pages 223–239. Springer, 2010.
- [14] Tamara L Berg, Alexander C Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *Proceedings of European Conference on Computer Vision*, pages 663–676. 2010.
- [15] Thomas Berg and Peter N Belhumeur. Tom-vs-pete classifiers and identity-preserving alignment for face verification. In *Proceedings of the British Machine Vision Conference*, volume 2, page 7, 2012.

- [16] Thomas Berg and Peter N Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 955–962, 2013.
- [17] Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2011–2018, 2014.
- [18] Lukas Bossard, Matthias Dantone, Christian Leistner, Christian Wengert, Till Quack, and Luc Van Gool. Apparel classification with style. In *Proceedings of Asian Conference on Computer Vision*, pages 321–335. Springer, 2013.
- [19] Léon Bottou. Stochastic gradient tricks. In *Neural Networks, Tricks of the Trade, Reloaded*, Lecture Notes in Computer Science (LNCS 7700), pages 430–445. 2012.
- [20] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the International Conference on Machine Learning*, pages 129–136, 2007.
- [21] R Caruna. Multitask learning: A knowledge-based source of inductive bias. In *Machine Learning: Proceedings of the Tenth International Conference*, pages 41–48, 1993.
- [22] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.
- [23] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11:1109–1135, 2010.

-
- [24] Bor-Chun Chen, Yan-Ying Chen, Yin-Hsi Kuo, and Winston H Hsu. Scalable face image retrieval using attribute-enhanced sparse codewords. *IEEE Transactions on Multimedia*, 15(5):1163–1173, 2013.
- [25] David M Chen, Sam S Tsai, Vijay Chandrasekhar, Gabriel Takacs, Ramakrishna Vedantham, Radek Grzeszczuk, and Bernd Girod. Inverted index compression for scalable image matching. In *Proceedings of the Data Compression Conference*, page 525, 2010.
- [26] Huizhong Chen, Andrew Gallagher, and Bernd Girod. Describing clothing by semantic attributes. In *Proceedings of European Conference on Computer Vision*, pages 609–623, 2012.
- [27] Qiang Chen, Junshi Huang, Rogerio Feris, Lisa M Brown, Jian Dong, and Shuicheng Yan. Deep domain adaptation for describing people based on fine-grained clothing attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5315–5324, 2015.
- [28] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2016.
- [29] Ronan Collobert, Koray Kavukcuoglu, and Clément Farabet. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*, 2011.
- [30] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585, 2006.

-
- [31] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005.
- [32] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)*, 40(2):5, 2008.
- [33] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-theoretic metric learning. In *Proceedings of ACM International Conference on Machine Learning*, pages 209–216, 2007.
- [34] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Computer Vision and Pattern Recognition*, pages 248–255, 2009.
- [35] Wei Di, Catherine Wah, Anurag Bhardwaj, Robinson Piramuthu, and Neel Sundaresan. Style finder: Fine-grained clothing style detection and retrieval. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 8–13. IEEE, 2013.
- [36] Changxing Ding, Chang Xu, and Dacheng Tao. Multi-task pose-invariant face recognition. *IEEE Transactions on Image Processing*, 24(3):980–993, 2015.
- [37] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *Proceedings of International Conference on Machine Learning*, volume 32, pages 647–655, 2014.
- [38] Matthijs Douze, Arnau Ramisa, and Cordelia Schmid. Combining attributes and fisher vectors for efficient image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 745–752, 2011.

- [39] Max Ehrlich, Timothy J Shields, Timur Almaev, and Mohamed R Amer. Facial attributes classification using multi-task representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 47–55, 2016.
- [40] Mark Everingham, Andrew Zisserman, Christopher KI Williams, Luc Van Gool, Moray Allan, Christopher M Bishop, Olivier Chapelle, Navneet Dalal, Thomas Deselaers, Gyuri Dorkó, et al. The pascal visual object classes challenge results. 2007.
- [41] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1785. IEEE, 2009.
- [42] X Yu Felix, Rongrong Ji, Ming-Hen Tsai, Guangnan Ye, and Shih-Fu Chang. Weak attributes for large-scale image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2949–2956, 2012.
- [43] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.
- [44] Pedro F Felzenszwalb, Ross B Girshick, and David McAllester. Cascade object detection with deformable part models. In *Proceedings of the IEEE conference on Computer vision and pattern recognition*, pages 2241–2248, 2010.
- [45] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.
- [46] Vittorio Ferrari and Andrew Zisserman. Learning visual attributes. In *Advances in Neural Information Processing Systems*, pages 433–440, 2008.

- [47] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C Berg. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017.
- [48] Yanwei Fu, Timothy M Hospedales, Tao Xiang, and Shaogang Gong. Learning multi-modal latent attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2):303–316, 2014.
- [49] Xingyu Gao, Steven CH Hoi, Yongdong Zhang, Ji Wan, and Jintao Li. Soml: Sparse online metric learning with application to image retrieval. In *Association for the Advancement of Artificial Intelligence*, pages 1206–1212, 2014.
- [50] Ross Girshick. Fast r-cnn. In *Proceedings of International Conference on Computer Vision*, pages 1440–1448, 2015.
- [51] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.
- [52] Georgia Gkioxari, Bharath Hariharan, Ross Girshick, and Jitendra Malik. R-cnns for pose estimation and action detection. *arXiv preprint arXiv:1406.5212*, 2014.
- [53] Boqing Gong, Jianzhuang Liu, Xiaogang Wang, and Xiaoou Tang. 3d object retrieval with semantic attributes. In *Proceedings of the ACM International Conference on Multimedia*, pages 757–758, 2011.
- [54] Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *Proceedings of the European Conference on Computer Vision*, pages 392–407, 2014.

-
- [55] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.
- [56] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 2016.
- [57] M Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C Berg, and Tamara L Berg. Where to buy it: Matching street clothing photos in online shops. In *Proceedings of International Conference on Computer Vision*, pages 3343–3351, 2015.
- [58] Chris Harris and Mike Stephens. A combined corner and edge detector. In *Alvey vision conference*, volume 15, page 50. Citeseer, 1988.
- [59] Junfeng He, Jinyuan Feng, Xianglong Liu, Tao Cheng, Tai-Hsu Lin, Hyunjin Chung, and Shih-Fu Chang. Mobile product search with bag of hash bits and boundary reranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3005–3012, 2012.
- [60] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *Proceedings of the European Conference on Computer Vision*, pages 346–361, 2014.
- [61] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.
- [62] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.

- [63] Junshi Huang, Rogerio S Feris, Qiang Chen, and Shuicheng Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *Proceedings of International Conference on Computer Vision*, pages 1062–1070, 2015.
- [64] Junshi Huang, Si Liu, Junliang Xing, Tao Mei, and Shuicheng Yan. Circle & search: Attribute-aware shoe retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 11(1):3, 2014.
- [65] Ahmet Iscen, Giorgos Toliás, Philippe-Henri Gosselin, and Hervé Jégou. A comparison of dense region detectors for image search and fine-grained classification. *IEEE Transactions on Image Processing*, 24(8):2369–2381, 2015.
- [66] Anil K Jain and Aditya Vailaya. Image retrieval using color and shape. *Pattern Recognition*, 29(8):1233–1244, 1996.
- [67] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.
- [68] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. pages 304–317, 2008.
- [69] Herve Jegou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Perez, and Cordelia Schmid. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1704–1716, 2012.
- [70] Shuhui Jiang, Yue Wu, and Yun Fu. Deep bi-directional cross-triplet embedding for cross-domain clothing retrieval. In *Proceedings of the ACM on Multimedia Conference*, pages 52–56. ACM, 2016.

- [71] Alexis Joly and Olivier Buisson. Logo retrieval with a contrario visual query expansion. In *Proceedings of the ACM international conference on Multimedia*, pages 581–584, 2009.
- [72] Zhuoliang Kang, Kristen Grauman, and Fei Sha. Learning with whom to share in multi-task feature learning. In *Proceedings of the International Conference on Machine Learning*, pages 521–528, 2011.
- [73] M Hadi Kiapour, Kota Yamaguchi, Alexander C Berg, and Tamara L Berg. Hipster wars: Discovering elements of fashion styles. In *Proceedings of the European conference on computer vision*, pages 472–488. Springer, 2014.
- [74] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Image search with relative attribute feedback. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2973–2980, 2012.
- [75] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [76] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.
- [77] BG Kumar, Gustavo Carneiro, Ian Reid, et al. Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5385–5394, 2016.

- [78] Neeraj Kumar, Alexander Berg, Peter N Belhumeur, and Shree Nayar. Describable visual attributes for face verification and image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1962–1977, 2011.
- [79] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Attribute and simile classifiers for face verification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 365–372. IEEE, 2009.
- [80] Weicheng Kuo, Bharath Hariharan, and Jitendra Malik. Deepbox: Learning objectness with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2479–2487, 2015.
- [81] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.
- [82] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958, 2009.
- [83] Ke Li, Kaiyue Pang, Yi-Zhe Song, Timothy Hospedales, Honggang Zhang, and Yichuan Hu. Fine-grained sketch-based image retrieval: The role of part-aware attributes. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 1–9, 2016.
- [84] Ya Li, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Multi-task model and feature joint learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 3643–3649, 2015.
- [85] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017.

- [86] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755, 2014.
- [87] Hongye Liu, Yonghong Tian, Yaowei Yang, Lu Pang, and Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2167–2175, 2016.
- [88] Jiawei Liu, Zheng-Jun Zha, QI Tian, Dong Liu, Ting Yao, Qiang Ling, and Tao Mei. Multi-scale triplet cnn for person re-identification. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 192–196. ACM, 2016.
- [89] Jingen Liu, Benjamin Kuipers, and Silvio Savarese. Recognizing human actions by attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3337–3344, 2011.
- [90] Si Liu, Jiashi Feng, Zheng Song, Tianzhu Zhang, Hanqing Lu, Changsheng Xu, and Shuicheng Yan. Hi, magic closet, tell me what to wear! In *Proceedings of the ACM international conference on Multimedia*, pages 619–628. ACM, 2012.
- [91] Si Liu, Zheng Song, Guangcan Liu, Changsheng Xu, Hanqing Lu, and Shuicheng Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3330–3337, 2012.
- [92] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision*, pages 21–37, 2016.

- [93] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1096–1104, 2016.
- [94] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.
- [95] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
- [96] Ping Luo, Xiaogang Wang, and Xiaoou Tang. A deep sum-product architecture for robust facial attributes analysis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2864–2871, 2013.
- [97] Bangalore S Manjunath and Wei-Ying Ma. Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):837–842, 1996.
- [98] Jiri Matas, Ondrej Chum, Martin Urban, and Tomás Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, 2004.
- [99] Niall McLaughlin, Jesus Martinez del Rincon, and Paul C Miller. Person reidentification using deep convnets with multitask learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(3):525–539, 2017.
- [100] Masaru Mizuochi, Asako Kanezaki, and Tatsuya Harada. Clothing retrieval based on local similarity with multiple images. In *Proceedings of the ACM international conference on Multimedia*, pages 1165–1168, 2014.

-
- [101] Milind Naphade, John R Smith, Jelena Tesic, Shih-Fu Chang, Winston Hsu, Lyndon Kennedy, Alexander Hauptmann, and Jon Curtis. Large-scale concept ontology for multimedia. *IEEE Multimedia*, 13(3):86–91, 2006.
- [102] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.
- [103] Paul Over, George M Awad, Jon Fiscus, Brian Antonishek, Martial Michel, Alan F Smeaton, Wessel Kraaij, and Georges Quénot. Trecvid 2010—an overview of the goals, tasks, data, evaluation mechanisms, and metrics. 2011.
- [104] Makoto Ozeki and Takayuki Okatani. Understanding convolutional neural networks in terms of category-level attributes. In *Proceedings of the Asian Conference on Computer Vision*, pages 362–375. Springer, 2014.
- [105] Devi Parikh and Kristen Grauman. Relative attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 503–510, 2011.
- [106] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2751–2758, 2012.
- [107] Florent Perronnin, Yan Liu, Jorge Sánchez, and Hervé Poirier. Large-scale image retrieval with compressed fisher vectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3384–3391, 2010.
- [108] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. *Proceedings of the European Conference on Computer Vision*, pages 143–156, 2010.

- [109] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.
- [110] Pedro HO Pinheiro and Ronan Collobert. Recurrent convolutional neural networks for scene labeling. In *Proceedings of the International Conference on Machine Learning*, pages 82–90, 2014.
- [111] Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson. A baseline for visual instance retrieval with deep convolutional networks. *arXiv preprint arXiv:1412.6574*, 2014.
- [112] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.
- [113] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. *arXiv preprint arXiv:1612.08242*, 2016.
- [114] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.
- [115] Oren Rippel, Manohar Paluri, Piotr Dollar, and Lubomir Bourdev. Metric learning with adaptive density discrimination. *arXiv preprint arXiv:1511.05939*, 2015.
- [116] Marcus Rohrbach, Michaela Regneri, Mykhaylo Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele. Script data for attribute-based recognition of composite activities. In *Proceedings of the European Conference on Computer Vision*, pages 144–157, 2012.

- [117] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.
- [118] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
- [119] Olga Russakovsky and Li Fei-Fei. Attribute learning in large-scale datasets. In *Trends and Topics in Computer Vision*, pages 1–14. Springer, 2012.
- [120] Ruslan Salakhutdinov and Geoffrey E Hinton. Deep boltzmann machines. In *International Conference on Artificial Intelligence and Statistics*, volume 1, page 3, 2009.
- [121] Aleksandr Sayko and Anton Slesarev. Report on a baseline approach to the second msr-bing challenge on image retrieval, 2014.
- [122] Walter J Scheirer, Neeraj Kumar, Peter N Belhumeur, and Terrance E Boult. Multi-attribute spaces: Calibration for attribute fusion and similarity search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2933–2940. IEEE, 2012.
- [123] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.
- [124] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

-
- [125] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 806–813, 2014.
- [126] Christian Siagian and Laurent Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(2):300–312, 2007.
- [127] Behjat Siddiquie, Rogério Schmidt Feris, and Larry S Davis. Image ranking and retrieval based on multi-attribute queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 801–808, 2011.
- [128] E. Simo-Serra and H. Ishikawa. Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 298–307, 2016.
- [129] Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, and Francesc Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 118–126, 2015.
- [130] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.
- [131] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [132] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proceedings of the IEEE Conference on Computer Vision*, pages 1470–1477, 2003.

- [133] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, pages 1849–1857, 2016.
- [134] Fengyi Song, Xiaoyang Tan, and Songcan Chen. Exploiting relationship between attributes for improved face verification. *Computer Vision and Image Understanding*, 122:143–154, 2014.
- [135] Ran Tao, Arnold WM Smeulders, and Shih-Fu Chang. Attributes and categories for generic instance search from one example. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 177–186, 2015.
- [136] Lorenzo Torresani, Martin Szummer, and Andrew Fitzgibbon. Efficient object category recognition using classemes. *Proceedings of the European Conference on Computer Vision*, pages 776–789, 2010.
- [137] Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.
- [138] Andrea Vedaldi and Karel Lenc. Matconvnet: Convolutional neural networks for matlab. In *Proceedings of the ACM international conference on Multimedia*, pages 689–692, 2015.
- [139] Andrea Vedaldi, Siddharth Mahendran, Stavros Tsogkas, Subhransu Maji, Ross Girshick, Juho Kannala, Esa Rahtu, Iasonas Kokkinos, Matthew B Blaschko, David Weiss, et al. Understanding objects in detail with fine-grained attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3622–3629, 2014.
- [140] Fang Wang, Le Kang, and Yi Li. Sketch-based 3d shape retrieval using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1875–1883, 2015.

- [141] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1393, 2014.
- [142] Xi Wang, Zhenfeng Sun, Wenqiang Zhang, Yu Zhou, and Yu-Gang Jiang. Matching user photos to online products with robust deep features. In *Proceedings of the ACM on International Conference on Multimedia Retrieval*, pages 7–14, 2016.
- [143] Xianwang Wang and Tong Zhang. Clothes search in consumer photos via color matching and attribute learning. In *Proceedings of the ACM international conference on Multimedia*, pages 1353–1356. ACM, 2011.
- [144] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2015.
- [145] Yan Wang, Sheng Li, and Alex C Kot. Category-separating strategy for branded handbag recognition. In *International Symposium on Communications, Control and Signal Processing*, pages 61–64, 2014.
- [146] Yang Wang and Greg Mori. A discriminative latent model of object classes and attributes. In *Proceedings of the European Conference on Computer Vision*, pages 155–168, 2010.
- [147] Di Wei, Wah Catherine, Bhardwaj Anurag, Pira-muthu Robinson, and Sundaresan Neel. Style finder: fine-grained clothing style recognition and retrieval, 2013.
- [148] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb):207–244, 2009.

- [149] Lingxi Xie, Jingdong Wang, Bo Zhang, and Qi Tian. Fine-grained image search. *IEEE Transactions on Multimedia*, 17(5):636–647, 2015.
- [150] Kota Yamaguchi, M Hadi Kiapour, and Tamara L Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *Proceedings of International Conference on Computer Vision*, pages 3519–3526, 2013.
- [151] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg. Parsing clothing in fashion photographs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3570–3577, 2012.
- [152] Kota Yamaguchi, Takayuki Okatani, Kyoko Sudo, Kazuhiko Murasaki, and Yukinobu Taniguchi. Mix and match: joint model for clothing and attribute recognition. In *British Machine Vision Conf*, 2015.
- [153] Yang Yang, Shengcai Liao, Zhen Lei, and Stan Z Li. Large scale similarity learning using similar pairs for person verification. In *Association for the Advancement of Artificial Intelligence*, pages 3655–3661, 2016.
- [154] Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011.
- [155] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1331–1338, 2011.
- [156] A. Yu and K. Grauman. Fine-Grained Visual Comparisons with Local Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

-
- [157] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 192–199, 2014.
- [158] Felix X Yu, Liangliang Cao, Rogerio S Feris, John R Smith, and Shih-Fu Chang. Designing category-level attributes for discriminative visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 771–778, 2013.
- [159] Felix X Yu, Rongrong Ji, Ming-Hen Tsai, Guangnan Ye, and Shih-Fu Chang. Weak attributes for large-scale image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2949–2956, 2012.
- [160] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 799–807, 2016.
- [161] Wei Yu, HIT EDU, Kuiyuan Yang, COM Yalong Bai, Tianjun Xiao, and COM Hongxun Yao. Visualizing and comparing alexnet and vgg using deconvolutional layers.
- [162] Joe Yue-Hei Ng, Fan Yang, and Larry S Davis. Exploiting local features from deep networks for image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 53–61, 2015.
- [163] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4353–4361, 2015.
- [164] H. Zhan, B. Shi, and A. C. Kot. Cross-domain shoe retrieval with a semantic hierarchy of attribute classification network. *IEEE Transactions on Image Processing*, PP(99):1–1, 2017.

- [165] H. Zhan, B. Shi, and A. C. Kot. Fashion analysis with a subordinate attribute classification network. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 1506–1511, July 2017.
- [166] Huijing Zhan, Sheng Li, and Alex C Kot. Tagging the shoe images by semantic attributes. In *Proceedings of the IEEE International Conference on Digital Signal Processing*, pages 892–895, 2015.
- [167] Huijing Zhan, Boxin Shi, and Alex C. Kot. Cross-domain shoe retrieval using a three-level deep feature representation. In *International Symposium on Circuits and Systems*, 2017.
- [168] Huijing Zhan, Boxin Shi, and Alex C. Kot. Street-to-shop shoe retrieval. In *Proceedings of the British Machine Vision Conference*, 2017.
- [169] Huijing Zhan, Boxin Shi, and Alex C. Kot. Street-to-shop shoe retrieval with multi-scale viewpoint invariant triplet network. In *Proceedings of the IEEE International Conference on Image Processing*, 2017.
- [170] Cha Zhang and Zhengyou Zhang. Improving multiview face detection with multi-task deep convolutional neural networks. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 1036–1041. IEEE, 2014.
- [171] Ning Zhang, Ryan Farrell, Forrest Iandola, and Trevor Darrell. Deformable part descriptors for fine-grained recognition and attribute prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 729–736, 2013.
- [172] Ning Zhang, Manohar Paluri, Marc’Aurelio Ranzato, Trevor Darrell, and Lubomir Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1637–1644, 2014.

- [173] Xiaofan Zhang, Feng Zhou, Yuanqing Lin, and Shaoting Zhang. Embedding label structures for fine-grained feature representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1114–1123, 2016.
- [174] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*, pages 94–108. Springer, 2014.
- [175] Liang Zheng, Yi Yang, and Qi Tian. Sift meets cnn: A decade survey of instance retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.
- [176] Liang Zheng, Yali Zhao, Shengjin Wang, Jingdong Wang, and Qi Tian. Good practice in cnn feature transfer. *arXiv preprint arXiv:1604.00133*, 2016.
- [177] Zhun Zhong, Mingyi Lei, Shaozi Li, and Jianping Fan. Re-ranking object proposals for object detection in automatic driving. *arXiv preprint arXiv:1605.05904*, 2016.
- [178] Zhengzhong Zhou, Jingjin Zhou, and Liqing Zhang. Demand-adaptive clothing image retrieval using hybrid topic model. In *Proceedings of the ACM on Multimedia Conference*, pages 496–500, 2016.
- [179] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *Proceedings of the European Conference on Computer Vision*, pages 391–405, 2014.