Deep Domain Adaptation for Describing People Based on Fine-Grained Clothing Attributes

Qiang Chen[1*], Junshi Huang[3*], Rogerio Feris[2], Lisa M Brown[2], Jian Dong[3],Shuicheng Yan[3]

[1] IBM Research, Australia, [2] IBM T.J. Watson Research Center, [3] National University of Singapore

qiangchen@au1.ibm.com, {rsferis, lisabr}@us.ibm.com, {junshi.huang, a0068947, eleyans}@nus.edu.sg

Abstract

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

1. Introduction

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

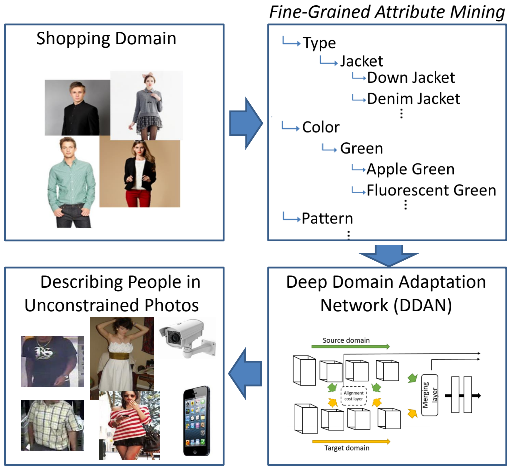

Figure 1. Overview of the proposed approach. We propose a novel deep domain adaptation method to bridge the gap between images crawled from online shopping stores and unconstrained photos. Another unique aspect of our work is the ability to decribe people based on their fine-grained clothing attributes.

Traditional computer vision methods for describing people, however, have only focused on a small set of coarse-grained attributes. As an example, the recent work of Zhang et al. [46] achieves impressive attribute prediction performance in unconstrained scenarios, but only considers nine human attributes. Existing systems for fashion analysis [3, 45] and people search in surveillance videos [11, 41] also rely on a relatively small set of clothing attributes.

Our work addresses the problem of describing people with very fine-grained clothing attributes. In particular, we consider attribute sub-categories that differ in subtle de-

[*] indicates equal contributions

[1] https://bankrobbers.fbi.gov/

tails, including many shades of clothing color (e.g., “Apple Green”, ”Fluorescent Green”, “Light Green”), different types of a particular garment (e.g., “Denim Jacket”, “Down Jacket”), and specific clothing patterns (e.g., “thin horizontal stripes”, “other types of stripes”). As far as we know, this is the first work to address this problem in a real scenario.

Directly tackling this problem is challenging because a large amount of annotated data is required to train such a large number of attribute models. In recent years, large-scale datasets such as ImageNet [6] and Labeled Faces in the Wild [21] have been built by leveraging vast amounts of visual data available on the web. However, most of the images obtained from online sources are either unlabelled or weakly labelled, often requiring costly manual annotation. In this work, we draw attention to e-Commerce websites, such as *Amazon.com* and *TMALL.com*, which contain *structured descriptions* of products and can be considered a reliable source of annotation. By leveraging this rich source of data from online shopping stores, we are able to collect a large-scale annotated dataset with around one million images along with fine-detailed attribute sub-categories.

Some of the typical images from these online shops are shown in Figure 1. As can be seen, there is a large discrepancy between these samples and the samples from our application domain, i.e., unconstrained photos captured by, for example, surveillance cameras or mobile phones. The online shopping images are often depicted with ideal lighting, standard pose, high resolution, and good quality, whereas these conditions cannot be guaranteed for images captured in the wild. Thus we investigate whether it is possible to perform domain adaptation to bridge the gap between these two domains.

We look into the newest weapon of computer vision research – deep learning approaches, which have been applied very effectively for visual recognition problems e.g., the large scale visual recognition ImageNet challenge [26] [6], and the object classification and detection tasks for PASCAL VOC datasets [13, 44]. Very recently, several works in computer vision have shown that it is generally effective to transfer a deep learning model learned from a large-scale corpus, e.g., ImageNet, to other tasks by using the activation maps of certain layers of Deep Convolutional Neural Networks (e.g., the second fully connected layer, FC2) [38, 20]. The underlying assumption of these methods is that the parameters of the low-level and mid-level network layers can be re-used across domains. As it may not be true for our domain adaptation problem, we aim to learn the domain-invariant hierarchical features directly, while transferring the domain information within intermediate layers. To this end, we design a specific double-path deep convolutional neural network for the domain adaptation problem. Each path receives one domain image as the input and they are connected through several alignment cost layers. These cost layers ensure that (1) the feature learning parameters for the two domains are not too far away and (2) similar labels encode similar high-level features.

Our **contributions** can be summarized as follows:

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

2. Related Work

Semantic visual attributes have received significant attention by the computer vision community in the past few years [29, 9, 37, 28]. Among other applications, attributes have been used for zero-shot classification [29], visual search [25, 40], fine-grained categorization [2], and sentence generation from images [27]. Most of these methods rely on costly manual annotation of labels for training the attribute models. Notable exceptions include techniques that mine attributes from web data [4], including sources such as Wikipedia [39] and online books [7]. Our work follows this direction, but we focus on cross-domain attribute mining, where the data is mined from online shopping stores and then adapted to unconstrained environments, using a novel deep domain adaptation approach.

Attribute Datasets. There are only a few attribute datasets with fine-grained annotations, for example, datasets related to detailed descriptions of birds [43] and aircrafts [42]. We push this envelope by proposing a new dataset of fine-grained clothing attributes. Compared to other clothing datasets for fashion analysis [3, 45], our proposed dataset has a much larger set of garments, including attribute *sub-categories* and a massive volume of training images per class.

Describing People by Attributes. Predicting human attributes [28, 1, 47, 11] is important for many surveillance applications, such as person re-identification across cameras [30], suspect search based on eyewitness testimonies [11, 41], and identification based on soft-biometrics [22]. Our approach deals with a *fine-grained* set of clothing attributes, which is around 10x larger than most previous methods, and requires minimal manual labeling for attribute learning.

Extracting clothing attributes for *analysis of fashion images* is another topic that has recently attracted interest [3, 45, 24, 33]. Previous methods developed for this application domain often focus on the clothing segmentation problem, considering pictures depicted in relatively simple poses, against relatively clean backgrounds. In our work, we study the domain adaptation problem from "clean" clothing images obtained from online shopping stores to images captured in unconstrained environments. Liu et al. [33] addressed a similar cross-domain clothing retrieval problem, but their work relies on a different methodology than ours, deals with a different application, and only considers a small set of coarse-grained attributes which are manually labeled.

Deep Learning. Deep Convolutional Neural Networks have recently achieved dramatic accuracy improvements in image classification [26], object detection [13], and many other computer vision areas, including attribute modeling [46, 34]. Recent improvements on deep learning include the use of drop-out [19] for preventing overfitting, more effective non-linear activation functions such as rectified linear units [14] or max-out [17], and richer modeling through Network-in-Network (NiN) [31]. In our work, we customize R-CNN and NiN for body detection, and propose a new deep domain adaptation approach to bridge the gap between the source and target clothing domain distributions.

Domain Adaptation. Many methods have been proposed for domain adaptation in visual recognition [16, 18, 15]. Recently, addressing this problem with deep neural networks has gained increased attention. The majority of existing approaches for domain adaptation or transfer learning with deep architectures rely on re-training the last few layers of the network using samples from the target domain, or instead performing *fine-tuning* of all layers using backpropagation at a lower learning rate [36, 38, 20]. However, these methods usually require a relatively large amount of training samples from the target domain to produce good results. In contrast, our method learns domain-invariant hierarchical features directly and transfers the domain information within intermediate layers, which we show to be much more effective. The work of Nguyen et al. [35] shares some of our motivations, but uses a different methodology based on dictionary learning.

A distinct method that was recently proposed for deep adaptation is DLID [5] which learns multiple unsupervised deep models directly on the source, target, and combined datasets, and uses a representation which is the concatenation of the outputs of each model as its adaptation approach. While this was shown to be an interesting approach, it is limited by its use of unsupervised deep structures, which have been unable to achieve the performance of supervised deep CNNs. Our method instead uses a supervised double path CNN with shared layers. It is able to leverage the extensive labeled data available in the source domain using a supervised model without requiring a significant amount of labeled target data.

3. Dataset Preparation

Although there are a few existing fashion datasets in the research community [45, 32], they are designed for the tasks of clothing parsing or human segmentation and no annotation of fine-grained attributes is included. Here we introduce our two sets of data and their statistics: (1) online shopping dataset obtained by crawling large amount of annotated images from online shopping stores and (2) "street" dataset which consists of both web street images and sample videos from surveillance cameras.

3.1. Online Shopping Dataset

3.1.1 Automatic Data Collection

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

3.1.2 Data Curation

In consumer photos or images from surveillance cameras, it might be difficult or impossible for a person to differentiate some attributes that could otherwise be discriminated in the online shopping domain. For instance, a security guard would likely not be able to tell the difference between a suspect wearing a "ginger yellow" shirt and another wearing a "turmeric" shirt. Therefore, we focus on a subset of the dataset mined from online shopping stores. We consider upper clothing only and select three attribute categories: type, color, and pattern. Several attributes are merged based on human perception. We also removed some attributes that are not well-defined, e.g., clothing images with "abstract patterns". Finally, we removed attributes for which the number of collected images is less than 1000. As a result, we focus on a subset of the data containing 341,021 images and 67 attributes, including 15 kinds of clothing types, 31 colors and 21 kinds of patterns. We denote this dataset as *Online-data*.

3.1.3 Building a Fine-grained Attribute Hierarchy

As described in section 3.1, the set of fine-grained attribute categories and labels mined from online shopping websites are given as a list of $< key, value >$ pairs without a hierarchical structure. We therefore organize this data into a semantic taxonomy of clothing attributes. We consider “type”, ‘color”, and ”pattern’ as the three higher-level categories. Each attribute is then classified into these three categories and further divided into semantic sub-categories. As an example, “wedding dress” and “sleeveless tank dress” are both sub-categories of “dress”, which is in turn a sub-category of “type”.

3.2. Street Dataset

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

Web street images. This data consists of two parts: (a) the standard Fashionate dataset [45], which consists of 685 street fashion images. This dataset has semantic segmentation labels (e.g., bags, pants), but no fine-grained clothing attributes. We fully annotated this dataset with our fine-grained attributes for evaluation purposes. We denote this data as *Street-data-a*. (b) the Parsing dataset [8] which consists of 8000 street fashion images. This dataset also has detailed segmentation labels, but we only used its images. We denote this data as *Street-data-b*.

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

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

4. Approach

We now introduce our solution to tackle the problem of describing people based on fine-grained attributes, as shown in Figure 3. First, we introduce an improved version of the R-CNN body detector which effectively localizes the clothing area. We then describe our proposed approach for attribute modeling and domain adaptation.

4.1. RCNN for Body Detection

Our body detection module is based on the R-CNN framework [13], with several enhancements made specifically for the clothing detection problem. It consists of three sub-modules. First, selective search is adopted to generate candidate region proposals. Then, a Network-in-Network (NIN) model is used to extract features for each candidate region. Finally, linear support vector regression (SVR) is used to predict the Intersection-over-Union (IoU) overlap of candidate patches with ground-truth bounding boxes. Next we introduce these components, and elaborate on the details of our enhancements.

Region Proposals. Due to the non-rigid property of clothing, standard selective search based on super-pixel proposal generation is shown to be more suitable to our detection task. Usually, about 2000 region proposals are generated per image. According to the clothing size in the training images, we discard the candidate regions with inappropriate size and aspect ratio. After that, about 100 hypotheses are left, which considerably reduces the number of noisy proposal regions and thus accelerates the feature extraction procedure.

Feature Extraction. The NIN model [31] is used to extract the high-level features from the candidate regions. Briefly, this model is pre-trained on the Imagenet Challenge dataset (ILSVRC-2014 classification task), and then fine-tuned using a subset of clothing images from our data.

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

In our implementation, we discretize the IoU values into ten intervals with a step of 0.1, and sample the equivalent training regions for each interval to balance the data during the training procedure. Lastly, the bounding box regression is employed to refine the selected proposal regions with the

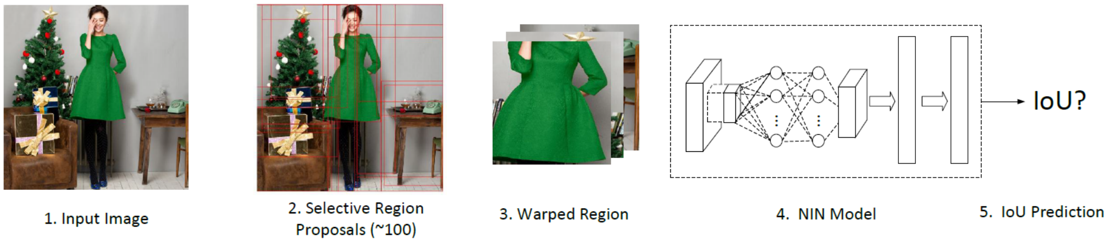

Figure 2. Enhanced R-CNN detection pipeline.

activation of the NIN fully-connected layer (FC2) as features.

4.2. Deep Domain Adaptation

Although we have collected a large scale dataset with fine-grained attributes, these images are taken in ideal pose/lighting/background conditions, so it is unreliable to directly use them as training data for attribute prediction in the domain of unconstrained images captured, for example, by mobile phones or surveillance cameras. In order to bridge this gap, we design a specific double-path deep convolutional neural network for the domain adaptation problem. Each path receives one domain image as the input, i.e., the street domain and the shop domain images. Each path consists of several convolutional layers which are stacked layer-by-layer and normally higher layers represent higher-level concept abstractions. Both of the two network paths share the same architecture, e.g., the same number of convolutional filters and number of middle layers. This way, the output of the same middle layer of the two paths can be directly compared. We further connect these paths through several alignment cost layers where the cost function is correlated with the similarity of the two input images. These alignment cost layers are included to ensure that (1) the feature learning parameters for the two domains are not too far away and (2) the high-level features have sufficient similarity along with the label consistency.

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

4.2.1 Alignment Layer Cost Function

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

$$f(s,t) = ||X_s - X_t|| \times \lambda\phi(s,t), \tag{1}$$

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

4.2.2 Discussion

It is worth noting the the following unique properties of the proposed DDAN: Consider a simplified CNN-based classification function, i.e. $y = f(g(x), w)$ where w are the classifier parameters (e.g., the final logistic regression layer) and $g(x)$ is the deep feature generator. In our domain adaptation problem, DDAN tries to align the target domain feature generator $g_{tgt}(x)$ with the source domain feature generator $g_{src}(x)$. As opposed to traditional domain adaptation approaches which try to align the features by finding a suitable subspace [16, 12], DDAN aims to align the high/middle level features directly during the training step of feature learning.

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

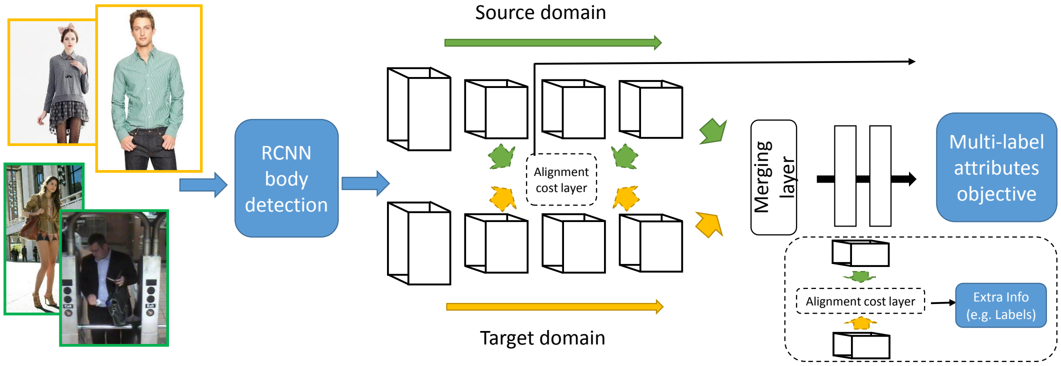

Figure 3. Overview of our proposed approach.

sistent output with the source domain feature so that we can directly apply new attribute classifiers to unseen labels learned from the source domain without additional cost.

Comparison with Siamese network: The structure of the proposed DDAN is similar to the Siamese network, often used for verification tasks. Both networks have two paths and a merging step. But the functions of these two key modules are quite different. The Siamese network calculates the difference of the two input channels and there is no back propagation channel to constrain the middle level representation.

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

5. Experiments

In this section, we present our implementation details, provide an extensive experimental analysis for evaluating our proposed approach, and showcase a couple of applications. We will be referring to the datasets described in section 3: *Online-data*, *Street-data-a*, *Street-data-b*, and *Street-data-c*.

5.1. Implementation Details

5.1.1 Network configuration

For each path of the DDAN, we configure the network with the same setting of the standard AlexNet [26], i.e. 5 convolutional layers and 3 fully connected layers, and the same filter numbers and neuron sizes. We have put the alignment cost layer at Conv5, FC1 and put the merging layer at FC2. During testing, we simply drop out the merging layer and use the target domain network. For the retrieval task, we used the FC2 output for both source and target domains.

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

Initialization: We tried out two initialization methods: (1) using random Gaussian values and (2) using the model learned from the ImageNet dataset as the initialization for both source and target domains. Generally, the second option gave us more stable results and fast convergence. Thus, we used this initialization method in the next sections. Normally 20 epochs are enough for the network training.

5.1.2 Learning setting

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

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

[2] code.google.com/p/cuda-convnet/

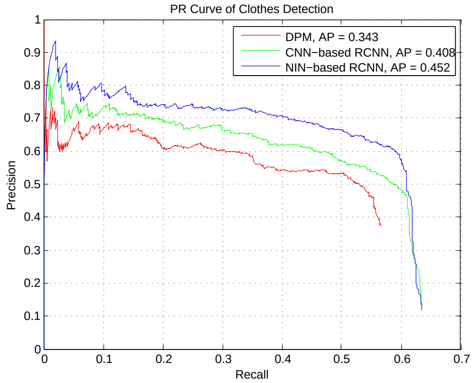

Figure 4. Precision-recall curves for body detection results on *Street-data-a* .

5.2. Exp1: RCNN body detection

For our body detection experiment, we annotated 4,000 images of the *Online-data* and 2,000 images of the *Street-data-b* with clothing bounding boxes for training. The *Street-data-a* dataset was used for validation. We compare the result of our enhanced RCNN-NIN detector with two baselines: (1) Deformable Part-based Model(DPM) [10] and (2) RCNN using a traditional CNN model [26] pre-trained on ImageNet [6] as feature generator. All of the baselines and the proposed method are tested using the same dataset.

As the performance of detection greatly affects the attribute learning, we set a strict evaluation metric for body detection. More specifically, we consider a detection to be correct only if the overlap of the prediction and ground-truth bounding boxes is over 0.6 instead of 0.5 as common in standard evaluation. We evaluate the performance of our body detector on *Street-data-a* with Precision Recall curves as shown in Figure 4. We also report the Average Precision (AP) result. As can be seen, our RCNN-NIN detector consistently outperforms the baselines (RCNN-NIN AP 0.452 vs DPM AP 0.343 and conventional-RCNN AP 0.408).

5.3. Exp2: Deep Domain Adaption for fine-grained attribute classification

We consider the following methods for comparison: (1) **CNN**, where we directly apply the model learned from the source domain to the target domain. (2) **CNN-FC2**, where we use the FC2 layer of a CNN model [26] as the features for training a classifier using street domain images. Since this method basically re-trains the attribute classifiers, it cannot predict unseen categories in the target domain. (3) **CNN-FT**, i.e., CNN fine-tuning, which keeps the original shop domain objective function, and then feeds the network with the street domain training samples.

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

Table 1. Fine-grained attributes classification results for *Street-data-c*.

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

Table 2. Fine-grained attributes classification results for *Street-data-a*.

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

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

Evaluation metrics: We used Top-1 (T1) and Top-1-base (T1-b) accuracy as the evaluation metrics, defined as follows. Top-1 accuracy is the standard evaluation metric for general classification problems. As claimed in Sec 3, we are working with a fine-grained attribute list. The attributes themselves naturally fall into a hierarchical structure. If the prediction and the groundtruth share the same immediate father node in the hierarchy, we consider it as a correct prediction. In this case, the accuracy we get is Top-1-base accuracy.

Result analysis: We present our results in Tables 1 and 2. For the *Street-data-c*, i.e., the surveillance video dataset, we only report the results of the fully unsupervised setting due to the lack of annotation for this domain. We can see consistent improvement of **DDAN-U** over directly applying the source domain CNN model with big margin. It shows that the feature learning can benefit from large amounts of unannotated target domain data. For the *Street-data-a* result, i.e., the street photo dataset result, we can see that our domain adaptation methods outperform the baselines **CNN**, **CNN-FC2** and **CNN-FT**. We can see that **DDAN-U** achieves the best results on most of the categories.

Overall, we notice that we can achieve much better results for "Type" and "Pattern" than "Color" categories, especially in the surveillance scenario. It is reasonable as

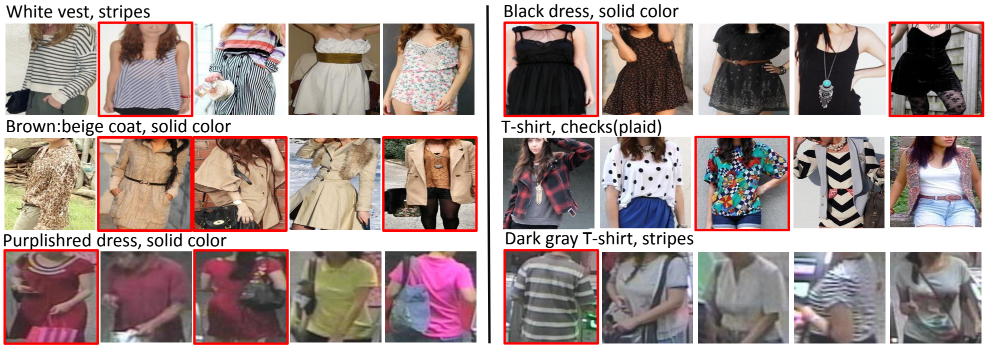
White vest, stripes
Black dress, solid color
Brown:beige coat, solid color
T-shirt, checks(plaid)
Purplishred dress, solid color
Dark gray T-shirt, stripes

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

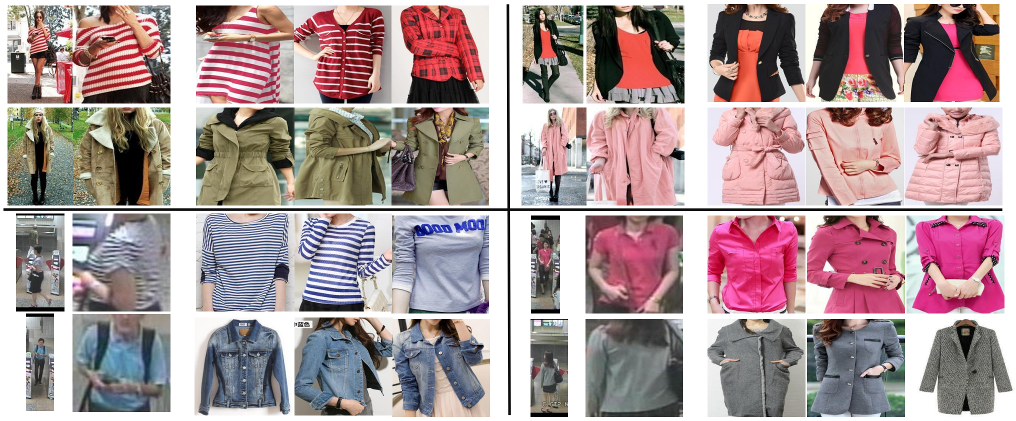

Figure 6. Application 2: Street2Shop clothing retrieval. Top 2 rows results are from *Street-data-a*, and the bottom two rows are from *Street-data-c*. We output the top 3 retrieval results for both datasets. Best viewed in original pdf file.

“Color” is very sensitive w.r.t. the lighting condition in the wild, while “Type” and “Pattern” are more related to the “shape” of the garments. Our domain adaptation framework reduces the gap between the two domains.

5.4. Application 1: Attribute-based people search

Here we showcase a few examples of attribute-based people search using the proposed system in Figure 5, e.g. finding people wearing a black-stripes T-shirt. We rank images based on the sum of the attribute confidence scores. We only show the top-5 ranked images due to space limitation. The images that exactly match the query are marked with red bounding box.

5.5. Application 2: Street2Shop clothing retrieval

As discussed in Sec 4.2, one major advantage of the proposed DDAN is that the output features share the same distribution of the source domain. So we can directly calculate the similarity of the two domain images without finding the common feature subspace or metric space. It provides great convenience for clothing retrieval – we can easily find the most similar online shopping clothes by looking at the feature similarity, e.g. the linear product distance. Some exemplar results are shown in Figure 6. We showcase the results for both *Street-data-a* and *Street-data-c* datasets. We output the top 3 retrieval results for both datasets.

6. Conclusion

In this paper, we presented a novel deep domain adaptation network for the problem of describing people based on fine-grained clothing attributes. As far as we know, this is the first work to address this problem in a real scenario. Our experiments show the advantage of the proposed approach over the baselines. We also showcased practical applications of this work. We are planning to make the full large-scale online shopping dataset available to the community, which in our opinion will be useful for various applications.

References

[1] L. Bourdev, S. Maji, and J. Malik. Describing People: A Poselet-Based Approach to Attribute Classification. In *ICCV*, 2011. 2

[2] S. Branson, C. Wah, B. Babenko, F. Schroff, P. Welinder, P. Perona, and S. Belongie. Visual Recognition with Humans in the Loop. In *ECCV*, 2010. 2

[3] H. Chen, A. Gallagher, and B. Girod. Describing Clothing by Semantic Attributes. In *ECCV*, 2012. 1, 2, 3

[4] X. Chen, A. Shrivastava, and A. Gupta. NEIL: Extracting Visual Knowledge from Web Data. In *ICCV*, 2013. 2

[5] S. Chopra, S. Balakrishnan, and R. Gopalan. Dlid: Deep learning for domain adaptation by interpolating between domains. In *ICML workshop on challenges in representation learning*, 2013. 3

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database . In *CVPR*, 2009. 2, 7

[7] S. Divvala, A. Farhadi, and C. Guestrin. Learning Everything about Anything: Webly-Supervised Visual Concept Learning. In *CVPR*, 2014. 2

[8] J. Dong, Q. Chen, X. Shen, J. Yang, and S. Yan. Towards unified human parsing and pose estimation. In *CVPR*, 2014. 4

[9] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing Objects by their Attributes. In *CVPR*, 2009. 2

[10] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008. 7

[11] R. Feris, R. Bobbitt, L. Brown, and S. Pankanti. Attribute-based people search: Lessons learnt from a practical surveillance system. In *ICMR*, 2014. 1, 2, 3

[12] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *ICCV*, 2013. 5

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv preprint arXiv:1311.2524*, 2013. 2, 3, 4

[14] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier networks. In *AISTATS*, 2011. 3

[15] B. Gong, F. Sha, and K. Grauman. Overcoming dataset bias: An unsupervised domain adaptation approach. In *NIPS Workshop on Large Scale Visual Recognition and Retrieval*, 2012. 3

[16] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012. 3, 5

[17] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio. Maxout networks. *arXiv preprint arXiv:1302.4389*, 2013. 3

[18] R. Gopalan, R. Li, and R. Chellappa. Unsupervised adaptation across domain shifts by generating intermediate data representations. *IEEE transactions on pattern analysis and machine intelligence*, 2014. 3

[19] G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Improving neural networks by preventing coadaptation of feature detectors. In *ArXiv e-prints*, 2012. 3

[20] J. Hoffman, E. Tzeng, J. Donahue, Y. Jia, K. Saenko, and T. Darrell. One-shot adaptation of supervised deep convolutional models. *CoRR*, abs/1312.6204, 2013. 2, 3

[21] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 2

[22] E. Jaha and M. Nixon. Soft Biometrics for Subject Identification using Clothing Attributes. In *IJCB*, 2014. 3

[23] K. Jarrett, K. Kavukcuoglu, M. Ranzato, Y. LeCun What is the Best Multi-Stage Architecture for Object Recognition? In *ICCV*, 2009.

[24] M. Kiapour, K. Yamaguchi, A. Berg, and T. Berg. Hipster Wars: Discovering Elements of Fashion Styles. In *ECCV*, 2014. 3

[25] A. Kovashka, D. Parikh, and K. Grauman. WhittleSearch: Image Search with Relative Attribute Feedback. In *CVPR*, 2012. 2

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 2, 3, 6, 7

[27] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. Berg, and T. Berg. Baby Talk: Understanding and Generating Simple Image Descriptions. In *ICCV*, 2011. 2

[28] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Attribute and Simile Classifiers for Face Verification. In *ICCV*, 2009. 2

[29] C. Lampert, H. Nickisch, and S. Harmeling. Learning To Detect Unseen Object Classes by Between-Class Attribute Transfer. In *CVPR*, 2009. 2

[30] R. Layne, T. Hospedales, and S. Gong. Person Re-Identification by Attributes. In *BMVC*, 2012. 2

[31] M. Lin, Q. Chen, and S. Yan. Network in network. *CoRR*, abs/1312.4400, 2013. 3, 4

[32] S. Liu, L. Liu, and S. Yan. Fashion Analysis: Current Techniques and Future Directions. *IEEE Multimedia*, 21(2):72–79, 2014. 3

[33] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *CVPR*, 2012. 3

[34] P. Luo, X. Wang, and X. Tang. A Deep Sum-Product Architecture for Robust Facial Attributes Analysis. In *ICCV*, 2013. 3

[35] H. V. Nguyen, H. T. Ho, V. M. Patel, and R. Chellappa. Joint hierarchical domain adaptation and feature learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, submitted*, 2013. 3

[36] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks. In *CVPR*, 2014. 3

[37] D. Parikh and K. Grauman. Relative Attributes. In *ICCV*, 2011. 2

[38] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. *arXiv preprint arXiv:1403.6382*, 2014. 2, 3

[39] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele. What Helps Where And Why? Semantic Relatedness for Knowledge Transfer. In *CVPR*, 2010. 2

[40] B. Siddiquie, R. S. Feris, and L. Davis. Image Ranking and Retrieval Based on Multi-Attribute Queries. In *CVPR*, 2011. 2

[41] D. A. Vaquero, R. S. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk. Attribute-based people search in surveillance environments. In *WACV*, 2009. 1, 3

[42] A. Vedaldi, S. Mahendran, S. Tsogkas, S. Maji, B. Girshick, J. Kannala, E. Rahtu, I. Kokkinos, M. Blaschko, D. Weiss, B. Taskar, K. Simonyan, N. Saphra, and S. Mohamed. Understanding Objects in Detail with Fine-grained Attributes. In *CVPR*, 2014. 2

[43] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. In *Technical Report*, 2011. 2

[44] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan. Cnn: Single-label to multi-label. *arXiv preprint arXiv:1406.5726*, 2014. 2

[45] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *CVPR*, 2012. 1, 2, 3, 4

[46] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. PANDA: Pose Aligned Networks for Deep Attribute Modeling. In *CVPR*, 2014. 1, 3

[47] J. Zhu, S. Liao, Z. Lei, D. Yi, and S. Li. Pedestrian attribute classification in surveillance: Database and evaluation. In *ICCV workshop on Large-Scale Video Search and Mining*, 2013. 2