Multi-Task curriculum transfer deep learning of clothing attributes
Dong, Q; Gong, S; Zhu, X; IEEE Winter Conference on Applications of Computer Vision

© 2017 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

For additional information about this publication click this link.
http://qmro.qmul.ac.uk/xmlui/handle/123456789/31283

Information about this research object was correct at the time of download; we occasionally make corrections to records, please therefore check the published record when citing. For more information contact scholarlycommunications@qmul.ac.uk
Multi-Task Curriculum Transfer Deep Learning of Clothing Attributes

Qi Dong, Shaogang Gong and Xiatian Zhu
School of EECS, Queen Mary University of London, UK
{q.dong, s.gong, xiatian.zhu}@qmul.ac.uk

Abstract

Recognising detailed clothing characteristics (fine-grained attributes) in unconstrained images of people in-the-wild is a challenging task for computer vision, especially when there is only limited training data from the wild whilst most data available for model learning are captured in well-controlled environments using fashion models (well lit, no background clutter, frontal view, high-resolution). In this work, we develop a deep learning framework capable of model transfer learning from well-controlled shop clothing images collected from web retailers to in-the-wild images from the street. Specifically, we formulate a novel Multi-Task Curriculum Transfer (MTCT) deep learning method to explore multiple sources of different types of web annotations with multi-labelled fine-grained attributes. Our multi-task loss function is designed to extract more discriminative representations in training by jointly learning all attributes, and our curriculum strategy exploits the staged easy-to-hard transfer learning motivated by cognitive studies. We demonstrate the advantages of the MTCT model over the state-of-the-art methods on the X-Domain benchmark, a large scale clothing attribute dataset. Moreover, we show that the MTCT model has a notable advantage over contemporary models when the training data size is small.

1. Introduction

Automatic recognition of clothing attributes in images from the wild, e.g. street views, has many applications from retail shopping to internet search and visual surveillance [23, 16]. However, clothing attribute recognition in-the-wild is challenging due to poor lighting, cluttered scenes, unknown viewpoint, and lacking image details (Figure 1 (b)). Deep learning exploits a large collection of imagery data from diverse sources, and has been shown to be very effective for image classification tasks [50, 48, 32, 3]. However, training a deep model requires extensive labelled information on the imagery data mostly generated by exhaustive manual annotation. For clothing attribute, the available labelled image data size is small and the number of fine-grained attribute categories is also limited [16, 54, 36].

To overcome the lack of manually labelled training data, web data mining provides a solution [12, 8, 27], from which a large number of web images with their meta-data can be collected without exhaustive manual labelling. For clothing, there is potentially a rich source of web images and their meta-text provided by online shops, where these meta-text contain fine-grained clothing attributes [8, 27, 40] (Figure 3). A notable characteristics of these online shopping clothing images is that they are well-posed by models captured against clean background in good lighting. A key challenge is how to transfer models trained using these clean shop images to recognise attributes in images captured in-the-wild from the streets, known as the domain drift problem [51, 41], where clean shop photos are considered as the source data from a source domain whilst unconstrained images from the wild are the target data from a target domain.
This work proposes a novel deep learning approach to modelling clothing attributes given a large sized \textit{source data} and very small sized \textit{target data} for model training. The model solves the challenging problem of transfer learning between the source and target domains when both source and target training data are weakly labelled at the image level. Our contributions are three-fold: (1) We formulate a novel \textit{Multi-Task Curriculum Transfer (MTCT)} deep learning approach to modelling clothing attributes. (2) In contrast to existing methods \cite{8, 27, 40}, which are limited in exploiting cross-domain data, the proposed MTCT deep attribute model is characterised by a multi-task joint learning deep network architecture for capturing the underlying correlations between different attributes with shared feature representations. (3) We implement a novel curriculum transfer deep learning strategy that aims to explore knowledge about attributes for solving more effectively the highly non-convex optimisation problem in model learning. This curriculum transfer learning strategy is motivated by cognitive studies \cite{13, 43, 33, 5}, with a multi-staged learning principle focusing on a simple task first before increasing the learning difficulty level, reminiscent to the human learning strategy. To our knowledge, this is the first attempt of formulating a curriculum learning strategy for deep learning of attributes, although there was an early study on language modelling \cite{5}. This is in strong contrast to the current popular end-to-end learning strategy deployed in deep learning \cite{28, 50}. Our extensive comparative evaluation using the X-Domain benchmark dataset \cite{8} against three state-of-the-art deep learning models for clothing attribute recognition, including FashionNet \cite{40}, DARN \cite{27} and DDAN \cite{8}, demonstrates a clear advantage of the proposed MTCT deep attribute model. Moreover, we show that MTCT also has a notable advantage over the state-of-the-arts when the target domain training data size becomes small.

2. Related Work

Attributes. Visual attributes have been widely exploited in computer vision, e.g. zero-shot learning \cite{35, 19}, face analysis \cite{34}, pedestrian description \cite{10, 20}, person re-identification \cite{36}, visual search \cite{31, 49, 16}. These studies typically pre-define a small set of attributes and require expensive manual labelling of training data. Different datasets often do not have consistent labelling, limiting their scalability. Driven by the desire for large quantities of cheaply labelled images, there are studies to explore web data sources for collecting large scale imagery data that come with “free” corresponding meta-text annotations \cite{9, 12, 44}. However, this poses a new problem in that the meta-data labels of these web data are less accurate nor consistent when compared with human manually labelled attributes. Studies on clothing modelling have been focused extensively on clothing segmentation against typically clean background \cite{7, 30, 59, 50}. There have also been efforts on shop-clothing image categorisation and retrieval using traditional hand-crafted features (e.g. SIFT, HOG) \cite{6, 7, 18, 55}, and more recently deep learning based features \cite{27, 8, 29, 40}. Given the costs of labelling therefore a lack of large scale clothing attribute annotations from different sources (domains), cross-domain clothing attribute learning is a challenging problem and largely under-studied \cite{40, 27}.

Deep Transfer Learning. Transfer learning for domain adaptation is a well studied area \cite{22, 24, 17, 47}. More recently, deep learning models are shown to be more robust than conventional models against domain changes, mainly due to the high modelling capacity and the availability of large scale labelled training data. However, the domain drift problem remains unsolved i.e. the performance of deep models still degrades in a new domain \cite{25}. A common approach to deep transfer learning is fine-tuning, using target domain data, the higher layers (FC layers) of a pre-trained deep model from the source data \cite{21, 46}. This assumes the availability of a large number of target training data, which is mostly not the case. A number of deep transfer learning models have been proposed for generic image categorisation \cite{25, 26, 11, 52, 41}, and more recently for fine-grained clothing attribute learning \cite{8, 27, 40}. Specifically, Chen et al. \cite{8} proposed a Deep Domain Adaptation Network (DDAN) with two branches by assigning one branch to a specific domain and then introducing two cross-branch connected layers that can enforce a feature distance between cross-domain images according to their attribute relations. A further extension of the DDAN model was also proposed by Huang et al. \cite{27}, which consists of a Dual Attribute-aware Ranking Network (DARN) to additionally accommodate image-level cross-domain correspondence as well as hierarchical structural knowledge of attributes in each network branch. More recently, Liu et al. \cite{40} introduced a FashionNet to model simultaneously both local attribute-level and holistic image-level clothing representations with a strong requirement on manually labelled clothing landmarks, making it less scalable than both DDAN and DARN networks. Our new MTCT (Multi-Task Curriculum Transfer) network shares some common characteristics with DDAN and DARN but also with a few important differences and advantages: Unlike DDAN, that only considers attribute labels, our method additionally models image-level cross-domain image pair relations for more effective domain adaptation. This is similar to DARN. However, whilst our domain transfer learning exploits multi-task/attribute feature learning, DARN only utilises shared common fully-connected feature representations for all attributes (Figure 2(c)). In contrast to the FashionNet, our MTCT net exploits cross-domain attribute learning \textit{without} the need for extensive clothing landmarks, more scalable to
wider applications. Moreover, our MTCT network explores uniquely the curriculum transfer learning strategy for more effective deep model learning. Our extensive comparative evaluation validates the advantages of MTCT over DDAN [8], DARN [27], and the FashionNet [40].

3. Multi-Task Curriculum Transfer Network

3.1. Problem Definition

To construct a deep model capable of recognising fine-grained clothing attributes on images in-the-wild (target domain), we collect clothing images and their meta-label as attributes \( \{ z_i \}_{i=1}^{n_{\text{attr}}} \) (e.g. clothing category, collar style) automatically from a range of online shopping web-sites, with a total of \( n_{\text{attr}} \) different attribute categories, each category \( z_i \) having its respective value range \( Z_i \). Intrinsically, this is a multi-label recognition problem since the \( n_{\text{attr}} \) attribute co-exist in every clothing image and may be assigned to different values.

Suppose (1) we have a collection of \( n_t \) target training images \( \{ I_t \}_{t=1}^{n_t} \) along with their attribute annotation vectors \( \{ a_{t}^i \}_{i=1}^{n_{\text{attr}}} \), and \( a_{t,j}^i \) refers to the \( j \)-th attribute value of target image \( I_t \); there are also \( n_s \) source training images \( \{ I_s \}_{s=1}^{n_s} \) with corresponding attribute vectors \( \{ a_{s}^i \}_{i=1}^{n_{\text{attr}}} \); \( n_s >> n_t \), that is, the number of labelled source images is much greater than that of labelled target images. Moreover, (2) we have access to \( n_{\text{pw}} \) pair correspondences between target and source clothing images, e.g. selfie images taken by shopping customers with known pairing to the online images of the same clothes (Figure 1). This cross-domain pair relation is useful in bridging the large domain gap by transferring attribute knowledge encoded in the source domain to the target domain with much less labelled data. It is worth noting that these two types of supervised learning lie at different levels:

Most attributes are localised to image regions, even though the location information is not provided in the annotation. Cross-domain pair labels are at the holistic image-level. We consider this not only a multi-label learning problem – joint learning for mutually correlated attribute labels, but also a multi-task transfer learning problem – inter-dependently learning the best individual attribute prediction given both local and holistic cross-domain annotations.

3.2. Network Architecture Design

Our MTCT deep model has two components: (I) multi-task deep learning, (II) curriculum transfer deep learning.

(I) Multi-Task Deep Learning. Clothing attributes co-occur selectively and to explore this inherent constraint for more reliable attribute prediction, we wish to model multi-attribute correlations by formulating a Multi-Task Network (MTN). This implements the multi-task learning principle [15]1 in a deep model. Although sharing a similar spirit of multi-task regression networks for face modelling [59]57, in this MTN model we learn a multi-task discriminative network for clothing modelling. Compared to independent attribute modelling, such multi-task learning also involves a smaller number of to-be-learned model parameters and thus with a lower model overfitting risk towards the given training data, beyond modelling mutual relations among different types of attributes and their common representations.

Specifically, the MTN consists of five stacked Network-In-Network (NIN) convolutional (conv) units [38] and \( n_{\text{attr}} \) parallel branches, with each branch representing a three layers of Fully-Connected (FC) sub-network for modelling one of the \( n_{\text{attr}} \) attributes respectively (Figure 2a,d)). The neuron number in the output-layer of \( i \)-th branch is \( |Z_i| \), i.e. the number of corresponding all possible attribute values \( a_i \). For model training, we utilise the Softmax loss function within any branch to model mutually exclusive relations among the attribute values for each attribute category by firstly predicting the \( j \)-th attribute posterior probability of image \( I_t \) over the ground truth \( a_{i,j} \):

\[
p(y_{i,j} = a_{i,j} | x_{i,j}) = \frac{\exp(W_{j}^T x_{i,j})}{\sum_{k=1}^{|Z_i|} \exp(W_{k}^T x_{i,j})}
\]

where \( x_{i,j} \) refers to the feature vector for \( j \)-th attribute, and \( W_{k} \) to the corresponding prediction function parameter, then computing the overall loss on a batch of \( n_{\text{bs}} \) images as the average additive summation of attribute-level loss with equal weight:

\[
l_{\text{sm}} = -\frac{1}{n_{\text{bs}}} \sum_{i=1}^{n_{\text{bs}}} \sum_{j=1}^{n_{\text{attr}}} \log \left( p(y_{i,j} = a_{i,j} | x_{i,j}) \right)
\]

This design above allows to jointly learn both attribute-generic (by all shared conv layers) and attribute-specific (by individual FC layer branches) discriminative features. In this context, each branch corresponds to a specific learning task responsible for the assigned attribute modelling.

The proposed MTN is similar to the DARN model [27] but with a crucial difference, that is, the attribute-specific branch in DARN contains only the last FC layer which serves as an attribute prediction function (Figure 2 (c)), therefore no attribute-specific representation learning in DARN and our experiments show this is less effective in learning discriminative features as compared to the proposed MTN, where FC1,2 layers are explicitly allocated for this purpose in each branch. As all clothing attributes are jointly modelled in DARN, we refer it as Joint Attribute Convolutional Neural Network (JAN) in the experimental evaluation reported in Section 4.

Learning the MTN model requires a large amount of training data1 whilst we usually only have very limited labelled target images. To overcome this problem, we want to

1Each MTN has 19 conv, 3 max-pooling and 27 (9 x 3) FC layers. The
explore richer source domain labelling information through cross-domain transfer learning.

**II Curriculum Transfer Learning.** To transfer source annotation knowledge to sparsely labelled target domain, we formulate a Curriculum Transfer (CT) learning strategy for deep learning. This is motivated by cognitive studies that suggest a better learning strategy adopted by human and animals is to start with learning easier tasks before gradually increasing the difficulties of the tasks, rather than to blindly learn randomly organised tasks [13, 43, 33].

In our cross-domain clothing attribute learning context, source and target domain data present naturally this easy-hard knowledge distribution: Source images taken from professional models are much easier to learn than target domain data, particularly from the lower layers [58, 28, 48, 46]. The easy-stage (Stage-1) of our CT learning strategy constructs a target MTN by sharing all attribute labels. (3) We create a target MTN by sharing all attributes labels. We train the whole MTN on source images with their attribute labels. (2) Stage-2: Learning to capture harder cross-domain knowledge by embedding cross-domain image pair information, and simultaneously appending harder target images into the model training process. As such, attribute labels of non-paired source images can also be exploited in addition to the cross-domain paired images. We instantiate this CT strategy by formulating a MTN based deep architecture.

**CT Learning Stage-1: Easy Model Learning and Transfer.** The main purpose of this first CT learning stage is to extract well-defined attribute features from large quantities of easy-to-learn source images and then directly transfer to the target domain. The intuition is that, the high non-convex model optimisation problem we need to address can be made not only simpler but also benefiting from better identified local minima for subsequent incremental learning if the starting sub-tasks are restricted to less hard tasks [5]. This principle is consistent with the notion of adaptive value [13]. This direct feature transfer from source to target domain exploits the characteristics of deep learning features being capable of capturing hierarchical information and independent of the training data, particularly from the lower layers [58, 28, 48, 46].

The easy-stage (Stage-1) of our CT learning strategy constructs a source and a target MTN model as follows: (1) Following common practice [50], we pre-train all NIN layers of a source MTN using the training data of ImageNet-1K [45] for obtaining a good parameter initialisation. (2) We train the whole MTN on source images with their attribute labels. (3) We create a target MTN by sharing all

MTCT model has in total 79.4 million parameters (57 million required fine-tuning). In comparison, DARN [27] and FashionNet [40] have 73 million and 135 million parameters respectively (all required fine-tuning).
parameters from the source MTN. In this way, the attribute information is transferred from source to target domain.

**CT Learning Stage-2: Hard Model Learning and Transfer.** We build on Stage-1 to transfer harder cross-domain pair relational knowledge and perform incremental learning on harder target data. This is achieved by constructing a three-stream MTN (3MTN) architecture consisting of two identical copies of the source MTN network and the target MTN obtained from Stage-1 (Figure 2(b)), taking as input cross-domain image triplets in the form of “\{Target \(I_t\), Positive Source \(I_{ps}\), Negative Source \(I_{ns}\)\)” where Target \(I_t\) and Positive Source \(I_{ps}\) are of the same clothing (obtained from the cross-domain pairing labelling), whereas Target \(I_t\) and Negative Source \(I_{ns}\) are of different clothing items. We require that feature similarity value between Target and Positive Source is greater than that between Target and Negative Source simultaneously. To this end, we consider the learning-to-rank approach to model optimisation and exploit the t-distribution Stochastic Triplet Embedding (t-STE) loss function due to its strength in discovering underlying data structure [53]:

\[
l_{t-STE} = \sum_{(I, I_{ps}, I_{ns}) \in T} \log \left( \frac{1 + \frac{\|f_t(I_t) - f_{ps}(I_{ps})\|_2^2}{\alpha \|f_t(I_t) - f_{ns}(I_{ns})\|_2^2}}{1 + \frac{\|f_t(I_t) - f_{ns}(I_{ns})\|_2^2}{\alpha}} \right)^\beta
\]

where \(\alpha\) denotes the freedom degree of the Student kernel; \(\beta = -\frac{1}{(1+\alpha)^2}\); \(f_t(\cdot)\) and \(f_{ps}(\cdot)\) refer to the feature extraction function for the target and source MTN respectively, that is, the vectorised feature maps of the conv5 layer used as the sample feature in each stream (Figure 2(b)).

Concurrently, we learn all FC layers of each attribute-specific branch in the Target stream with the Softmax loss for obtaining the final attribute recognition model (Figure 2(b)). In practice, we found that fine-tuning FC layers in the source stream helps due to the mutual benefits between the two domains. As a result, all layers are frozen except conv5 of the Target MTN stream and all FC layers of both streams during the CT Stage-2 learning.

**Clothing Attribute Recognition:** Our learning aim is to obtain an optimised target MTN model for attribute recognition in-the-wild. This is achieved during model training by learning a source MTN for extracting and transferring localised attribute information, followed by optimising a 3MTN for transferring cross-domain pairing knowledge and adapting the target MTN to data from the wild. During model deployment, we solely utilise the target MTN for fine-grained clothing attribute recognition on unconstrained images. In the next section, we shall demonstrate the effectiveness of the proposed model when compared against the state-of-the-arts.

4. Experiments

4.1. Dataset and Evaluation Protocol

We utilised the Cross-Domain (X-Domain) clothing attribute dataset [3] for our comparative evaluation. Specifically, this X-Domain dataset contains two different image source domains: (1) The shop domain, online stores such as Amazon.com and TMall.com; (2) The street domain where consumer images are available.

Specifically, there are 245, 467 shop images each associated with web meta-data including \(\leq 9\) attribute/value pairs. These nine fine-grained clothing attributes are: category, button, colour, length, pattern, shape, collar, sleeve-length (slv-len), sleeve-shape (slv-shp). There may be varying numbers of optional values for different attributes, ranging from 6 (slv-len) to 55 (colour) and a total of 178 distinct values over all attributes. Therefore, these clothing attributes are rather fine-grained, possibly with subtle visual appearance dissimilarity between different attribute values, e.g. Woollen-Coat versus Cotton-Coat. Note that these attribute data were webly annotated at the image-level and thus weakly-supervised with no specified attribute location.

We also have 46, 769 street images from customer reviews of a proportion of shop image webpages. Among these 46, 769 in-the-wild images, there are 14, 186 cross-domain pairing with the shop images. The remaining 231, 281 (245, 467 – 14, 186) shop images are non-paired. In our evaluations, we consider the shop and street domains as the source and target domains, respectively.

**Evaluation Protocol.** On our copy of the X-Domain dataset, we performed the following data partition for cross-domain attribute recognition evaluation. For the shop domain, we randomly selected 165, 467 images as training data and the remaining 80, 000 as test images. For the street domain, 36, 769 were randomly selected for training and 10, 000 for test.

For quantitative evaluation, we adopted both per-class (i.e. per-attribute) and per-instance (i.e. per-image) based metrics. For the former, we used Average Precision (AP\(^{cls}\)) for each attribute class and mean Average Precision (mAP\(^{cls}\)) over all classes [3]. For the latter, we first computed per-image attribute Precision and Recall, then averaged both over all images to obtain mean Precision (mP\(^{ins}\)) and Recall (mR\(^{ins}\)) [40].

4.2. Implementational Considerations

**Clothing Detection.** As input images are not accurately cropped, clothing detection is necessary for reducing the negative impact of background clutter. We performed clothing detection by a customised Faster R-CNN model

---

*In our experiments, we collected 100% shop images of the X-Domain dataset, but only 69% of the cross-domain pairing images were available from the X-Domain URLs given by [3] due to commercial copyrights.*
Table 1. Comparing state-of-the-art clothing attribute recognition methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Category</th>
<th>Button</th>
<th>Colour</th>
<th>Length</th>
<th>Pattern</th>
<th>Shape</th>
<th>Collar</th>
<th>Slv-Len</th>
<th>Slv-Shp</th>
<th>mAP^ins</th>
<th>mPpos</th>
<th>mRpos</th>
</tr>
</thead>
<tbody>
<tr>
<td>DDAN [8]</td>
<td>12.56</td>
<td>24.13</td>
<td>20.72</td>
<td>35.91</td>
<td>61.67</td>
<td>47.14</td>
<td>31.17</td>
<td>80.63</td>
<td>73.96</td>
<td>43.10</td>
<td>45.41</td>
<td>52.20</td>
</tr>
<tr>
<td>DARN [27]</td>
<td>32.55</td>
<td>37.48</td>
<td>58.24</td>
<td>51.49</td>
<td>67.53</td>
<td>47.70</td>
<td>47.77</td>
<td>82.04</td>
<td>73.72</td>
<td>53.61</td>
<td>57.79</td>
<td>67.29</td>
</tr>
<tr>
<td>FashionNet [40]</td>
<td>55.85</td>
<td>39.52</td>
<td>60.33</td>
<td>53.08</td>
<td>68.65</td>
<td>49.79</td>
<td>52.17</td>
<td>83.79</td>
<td>75.34</td>
<td>59.84</td>
<td>59.97</td>
<td>69.74</td>
</tr>
<tr>
<td>MTCT</td>
<td>65.96</td>
<td>43.57</td>
<td>66.86</td>
<td>58.27</td>
<td>78.55</td>
<td>51.40</td>
<td>58.97</td>
<td>86.05</td>
<td>77.54</td>
<td>64.35</td>
<td>64.97</td>
<td>75.66</td>
</tr>
</tbody>
</table>

Specifically, we first trained our detector on PASCAL VOC2007 training data [14] followed by fine-tuning on an assembled clothing dataset consisting of 8,000 street/shop photos (with box annotation available) from [29] and 4,000 fashion images (with boxes generated from available pixel-level labels) from [37].

Parameter Settings. For training the MTCT, the momentum was set at 0.9 and weight-decay at 0.0005, same as in NIN [38] and AlexNet [32]. The batchsize was set at 256 limited by the GPU memory size. The learning rates were set empirically, by the training loss change, at 0.001 for pre-training on the ImageNet, and 0.0001 for fine-tuning on the source/target clothing data. For training the other compared models, we used the same parameter settings given by the authors, otherwise same as for MTCT.

4.3. Evaluation Choices

We compared our MTCT deep attribute model with 3 state-of-the-art models and 4 different variants of our model design: (1) Deep Domain Adaptation Network (DDAN) [8]: A cross-domain attribute recognition model capable of learning domain invariant features by particularly aligning middle level representations of two domains during the training stage. (2) Dual Attribute-aware Ranking Network (DARN) [27]: A domain adaptation deep model that is trained and optimised with both attribute annotations and cross-domain pair correspondences in an end-to-end learning manner. (3) FashionNet [40]: A very recent clothing analysis model specially designed for multiple recognition tasks such as attribute and landmark detection. We implemented this model excluding landmark detection branch since landmark labels are not available in this real-world X-Domain dataset. We also compared MTCT against four different MTN based models to evaluate the role of the individual components in the MTCT model design. These are: (4) No Adaptation (NoAdpt): We train a given network using the labelled source data and directly deploy it for the target text data. This simple scheme has shown power and superiorities in many applications [48] [46] [23] due to the great generality of deep features by benefitting from large scale diverse training data. (5) JAN (NoAdpt): we set DARN without adaptation as baseline, i.e. training JAN in the source domain and then directly testing it on the target domain. (6) United Domains (UD): We train a given model on the union of source and target training data. Compared with NoAdpt, more data are exploited for model optimisation so that the feature generality may be further improved. (7) Fine-Tune based Transfer (FTT): We first train a given model on the source training data, then fine-tune the fully-connected layers on the target data. This is the vanilla transfer learning method commonly adopted in the literature [58]. Finally, we have the (8) Multi-Task Curriculum Transfer (MTCT): Our full model exploiting both multi-task and curriculum transfer learning. For fair comparison, all methods have access to the same training data, learned with their designed optimisation algorithms, and evaluated on the same test set.

4.4. Comparison to the State-Of-The-Art

We evaluated comparatively MTCT model performance on clothing attribute recognition. The comparative results with state-of-the-arts are presented in Table 1. It is evident that the proposed MTCT achieves the best results under all evaluation metrics, e.g. outperforming the best alternative FashionNet by 4.51% in mAP^ins. This suggests the superiority of our method in extracting and transferring source annotation information into the sparsely labelled and challenging target domain. More specifically, we can draw the following observations. Firstly, DDAN is the worst performer among all competitors, mainly because this model is less effective in mining rich non-paired source images, e.g. optimised with cross-domain paired images whilst most non-paired ones are not selected for model learning. By jointly modelling data from both domains with additional pair relations, DARN is able to extract and transfer more source information. However, it also suffers from the same problem above as DDAN. FashionNet surpasses DARN by learning the union of both domains and exploiting a more powerful basis architecture VGG16 [50] which is stronger
than both the AlexNet [32] used by DDAN and NIN [38] used by DARN. Despite that, FashionNet is still inferior to the NIN based MTCT model due to the former’s higher model overfitting risk caused by much more parameters required learning (135 million of FashionNet vs. 57 million of MTCT parameters required fine-tuning in model learning) and the ignorance of domain discrepancy in learning strategy (end-to-end vs. curriculum staged learning).

4.5. Effects of Multi-Task and Transfer Learning

We evaluated the effectiveness of the multi-task and curriculum transfer learning components in the MTCT model (Table 2). By explicitly learning individual attribute representations, the MTN(NoAdpt) without curriculum transfer improves model generalisation over JAN(NoAdpt) (DARN [27] without transfer learning). This demonstrates the benefit of multi-task learning. When additional 3,769 labelled target images (vs. 245,467 source images) were utilised, MTN(UD) improves further attribute recognition accuracy over MTN(NoAdpt). This supports the general observation that learning from target domain data is beneficial when there is a large discrepancy between the source and target domains. Given a vanilla fine-tuning transfer learning MTN(FTT), model performance is further boosted, which confirms similar findings elsewhere [4, 2]. MTN(FTT) is a special case of Curriculum Learning in that the initialisation by source data is an easier learning task whilst fine-tuning on target data is a harder task, but without learning cross-domain pairing information.

Our MTCT model is a fusion of DARN and vanilla transfer learning MTN(FTT) with unique advantages over both: (1) Similar to DARN, MTCT exploits cross-domain pairing information in model optimisation but with a lot less parameters (57 million vs. 73 million of DARN); also different from DARN in that MTCT adopts curriculum learning rather than end-to-end, first learning from easier attribute labels (source) then from harder pairing relations. (2) Similar to FTT, MTCT is optimised in a staged process, first learning easier source data then harder target data, plus harder still pairing data. Therefore, MTCT model explores two curriculum learning criteria, one on training data selection and another on supervision label difficulties. Qualitative evaluation is shown in Figure 3, where success cases showing the robustness of MTCT against cluttered background and complex viewing conditions. Figure 4 shows some failure cases under extreme poses and very challenging background ambiguities.

![Figure 3](image-url)  
Figure 3. A qualitative evaluation of our proposed MTCT method on unconstrained consumer images.

![Figure 4](image-url)  
Figure 4. Model robustness vs. target training data size.
4.6. Effects of Cross-Domain Training Data Size

We evaluated the robustness of different models against target training data size variation. For this evaluation, we reduced the number of target training images to \{75\%, 50\%, 10\%\} of the full training set and show respective results in Figure 5. It is evident that the proposed MTCT outperforms all competitors over different sparseness ratios. This demonstrates the advantages and scalability of our approach over alternatives models. Specifically, JAN(NoAdpt) utilises no target data so remains at just above 50\% mAP at a constant (the green dot on the right hand side above 0\%). DDAN stays at a low 42 \(-\) 45\% with little change, suggesting that transfer learning is difficult without exploiting cross-domain pair relations. As expected, the three models which have benefited from cross-domain pairing information all degrade with fewer training data available. Importantly, the MTCT model surpasses significantly other two models with 8.4\% relative improvement over the FashionNet, given only 3,676 labelled target images.

4.7. Further Analysis

(1) Automatic Clothing Detection. We evaluated the performance of our customised Faster R-CNN clothing detector. For this evaluation, we manually labelled clothing boxes on 400 images from X-Domain, including 200 shop and 200 consumer images. We set the correct detection Intersection over Union (IoU) threshold to 0.6. Our detector achieves 90.8\% recall on shop images and 71.2\% on in-the-wild images. This provides a more realistic testing platform for attribute recognition. Qualitative examples of clothing attribute detection and recognition, failure cases, and cross-domain clothing matching by attributes (red boxes) are shown in Figures 3, 4, and 6 respectively.

(2) Triplet Ranking vs. t-STE loss. Table 4 compares the performance of t-STE loss \[53\] against the common triplet ranking loss \[27\] in our MTCT network, showing that the t-STE loss function yields mAP 1.75\% performance advantage over the commonly used triplet ranking loss.

(3) Curriculum vs. End-to-End Transfer Learning. We evaluated the effectiveness of our CT method by comparing it with the popular End-to-End counterpart, both using our MTN architecture. This End-to-End baseline can be considered as an improved DARN model, i.e. replacing the JAN component of DARN with MTN. The results are shown in Table 3. It is clear that the proposed curriculum transfer is superior to end-to-end transfer learning, an improvement of 2.05\% in mAP\cls, suggesting that staged learning can better regularise deep model optimisation towards more discriminative local minima in the parameter space.

5. Conclusion

In this work, we formulated a Multi-Task Curriculum Transfer (MTCT) deep learning method for modelling fine-grained clothing attributes. We demonstrated its effectiveness in attribute recognition given unconstrained images taken from-the-wild (street views). This MTCT model (with 79.4 million parameters) outperforms the state-of-the-art FashionNet (with 135 million parameters) by 4.51\% in mAP\cls on the X-Domain benchmark. The proposed MTCT model is designed to optimise information transfer learning given large quantities of labelled information in a clean source domain and small sized labelled data in a noisy target domain in-the-wild. Specifically, MTCT exploits both a multi-task attribute learning deep network (MTN) and a staged curriculum learning strategy to maximise model learning. Moreover, we show the advantages of the MTCT over alternative models given decreased sizes of labelled target domain data, surpassing the FashionNet in performance on the X-Domain benchmark by \~8\% when only <4,000 target training images are available.
References


