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].

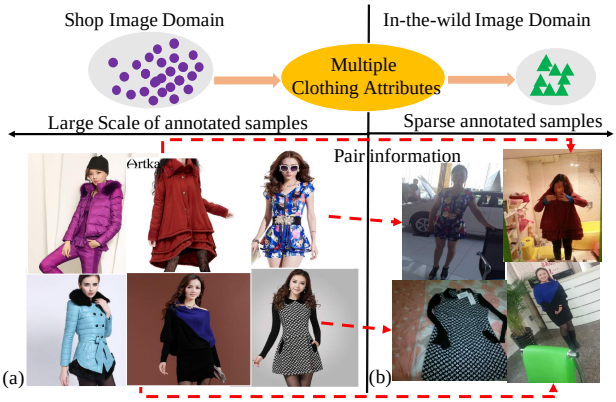


Figure 1. Clothing images of (a) professional models in shops and their corresponding instances (b) in-the-wild from the streets, with significant changes in appearance and background clutter.

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 *source data* and very small sized *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 *Multi-Task Curriculum Transfer* (MTCT) deep learning approach to modelling clothing attributes. (2) In contrast to existing methods [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 [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 [5]. This is in strong contrast to the current popular end-to-end learning strategy deployed in deep learning [28, 50]. Our extensive comparative evaluation using the X-Domain benchmark dataset [8] against three state-of-the-art deep learning models for clothing attribute recognition, including FashionNet [40], DARN [27] and DDAN [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 [35, 19], face analysis [34], pedestrian description [10, 20], person re-identification [36], visual search [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 [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 [7, 30, 39, 56]. There have also been efforts on shop-clothing image categorisation and retrieval using traditional hand-crafted features (e.g. SIFT, HOG) [6, 7, 7, 18, 55], and more recently deep learning based features [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 [40, 27].

Deep Transfer Learning. Transfer learning for domain adaptation is a well studied area [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 [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 [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 [25, 26, 11, 52, 41], and more recently for fine-grained clothing attribute learning [8, 27, 40]. Specifically, Chen et al. [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. [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. [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 *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_{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_{attr} attribute categories 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 $\{\boldsymbol{I}_i^t\}_{i=1}^{n_t}$ along with their attribute annotation vectors $\{\boldsymbol{a}_i^t\}_{i=1}^{n_t}$, and $\boldsymbol{a}_i^t = [a_{i,1}^t, \ldots, a_{i,j}^t, \ldots, a_{i,n_{\text{attr}}}^t]$ where $a_{i,j}^t$ refers to the j-th attribute value of target image $\boldsymbol{I}_i^t$; there are also n_s *source* training images $\{\boldsymbol{I}_i^s\}_{i=1}^{n_s}$ with corresponding attribute vectors $\{\boldsymbol{a}_i^s\}_{i=1}^{n_s}$; $n_s >> n_t$, that is, the number of labelled source images is much greater than that of labelled target images. Moreover, (2) we have access to n_{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_{attr} parallel branches, with each branch representing a three layers of Fully-Connected (FC) sub-network for modelling one of the n_{attr} attributes respectively (Figure 2(a,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 $\boldsymbol{I}_i$ over the ground truth $a_{i,j}$:

$$p(y_{i,j} = a_{i,j}|\boldsymbol{x}_{i,j}) = \frac{\exp(\boldsymbol{W}_j^\top \boldsymbol{x}_{i,j})}{\sum_{k=1}^{|Z_j|} \exp(\boldsymbol{W}_k^\top \boldsymbol{x}_{i,j})} \quad (1)$$

where $\boldsymbol{x}_{i,j}$ refers to the feature vector for j-th attribute, and $\boldsymbol{W}_k$ to the corresponding prediction function parameter, then computing the overall loss on a batch of n_{bs} images as the average additive summation of attribute-level loss with equal weight:

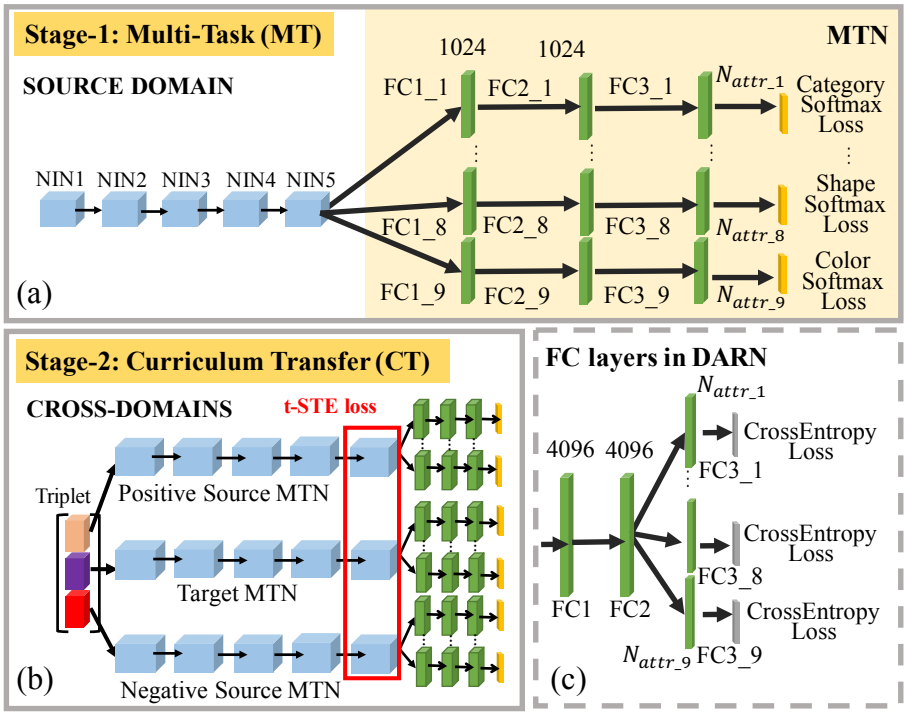
$$l_{\text{sm}} = -\frac{1}{n_{\text{bs}}} \sum_{i=1}^{n_{\text{bs}}} \sum_{j=1}^{n_{\text{attr}}} \log\Big(p(y_{i,j} = a_{i,j}|\boldsymbol{x}_{i,j})\Big) \quad (2)$$

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_3 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 $\text{FC}_{1,2}$ 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 data[1], whilst we usually only have very limited labelled target images. To overcome this problem, we want to

[1]Each MTN has 19 conv, 3 max-pooling and 27 (9×3) FC layers. The



(d) **MTN configurations**

	Layer Name	Parameters
NIN1	Conv1	7x7-96-2
	Conv1_1	1x1-96-1
	Conv1_2	1x1-96-1
	Maxpooling-3x3	
NIN2	Conv2	5x5-256-2
	Conv2_1	1x1-256-1
	Conv2_2	1x1-256-1
	Maxpooling-3x3	
NIN3	Conv3	3x3-512-1
	Conv3_1	1x1-512-1
	Conv3_2	1x1-512-1
NIN4	Conv4	3x3-1024-1
	Conv4_1	1x1-1024-1
	Conv4_2	1x1-512-1
	Conv4_3	1x1-384-1
NIN5	Conv5	3x3-512-2
	Conv5_1	1x1-512-1
	Conv5_2	1x1-512-1
	Maxpooling-3x3	
FC-1024 for any branch		
FC-1024 for any branch		
FC- N_{attr_i} for any branch		
Softmax for any branch		

Figure 2. (a)-(b): The MTCT network design, (c): The FC layers of DARN [27], (d): MTN configuration details.

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 images captured in-the-wild (Figure 1); Moreover, the difference in holistic (cross-domain pairing) and local annotations (source domain attributes) exhibits distinct degrees of learning complexity – attributes are localised thus specific whilst pair correspondences are holistic therefore abstract, with the latter has greater variations than the former.

Given these observations above, we propose a *two-stage curriculum transfer* (CT) learning strategy tailored for clothing attribute modelling: (1) Stage-1: Learning with clean (easier) source images and their attribute labels only; (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* of starting in developmental psychology [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* $\boldsymbol{I}_{\text{t}}$, *Positive Source* $\boldsymbol{I}_{\text{ps}}$, *Negative Source* $\boldsymbol{I}_{\text{ns}}$}" where *Target* $\boldsymbol{I}_{\text{t}}$ and *Positive Source* $\boldsymbol{I}_{\text{ps}}$ are of the same clothing (obtained from the cross-domain pairing labelling), whereas *Target* $\boldsymbol{I}_{\text{t}}$ and *Negative Source* $\boldsymbol{I}_{\text{ns}}$ 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]:

$$l_{\text{t-STE}} = \sum_{\{\boldsymbol{I}_{\text{t}},\boldsymbol{I}_{\text{ps}},\boldsymbol{I}_{\text{ns}}\}\in T} \log \frac{(1+\frac{\|f_{\text{t}}(\boldsymbol{I}_{\text{t}})-f_{\text{s}}(\boldsymbol{I}_{\text{ps}})\|^2}{\alpha})^{\beta}}{(1+\frac{\|f_{\text{t}}(\boldsymbol{I}_{\text{t}})-f_{\text{s}}(\boldsymbol{I}_{\text{ps}})\|^2}{\alpha})^{\beta}+(1+\frac{\|f_{\text{t}}(\boldsymbol{I}_{\text{t}})-f_{\text{s}}(\boldsymbol{I}_{\text{ns}})\|^2}{\alpha})^{\beta}} \quad (3)$$

where α denotes the freedom degree of the Student kernel; $\beta = -\frac{(1+\alpha)}{2}$; $f_{\text{t}}(\cdot)$ and $f_{\text{s}}(\cdot)$ 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 [8] for our comparative evaluations[2]. 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 ≤ 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 [8]. 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

[2]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 [8] due to commercial copyrights.

Table 1. Comparing state-of-the-art clothing attribute recognition methods.

Method	Category	Button	Colour	Length	Pattern	Shape	Collar	Slv-Len	Slv-Shp	mAP^{cls}	mP^{ins}	mR^{ins}
DDAN [8]	12.56	24.13	20.72	35.91	61.67	47.14	31.17	80.63	73.96	43.10	45.41	52.20
DARN [27]	52.55	37.48	58.24	51.49	67.53	47.70	47.77	82.04	73.72	57.61	57.79	67.29
FashionNet [40]	55.85	39.52	60.33	53.08	68.65	49.79	52.17	83.79	75.34	59.84	59.97	69.74
MTCT	**65.96**	**43.57**	**66.86**	**58.27**	**70.55**	**51.40**	**58.97**	**86.05**	**77.54**	**64.35**	**64.97**	**75.66**

[42]. 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 test data. This simple scheme has shown power and superiorities in many applications [48, 46, 28] due to the great generality of deep features by benefiting 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.

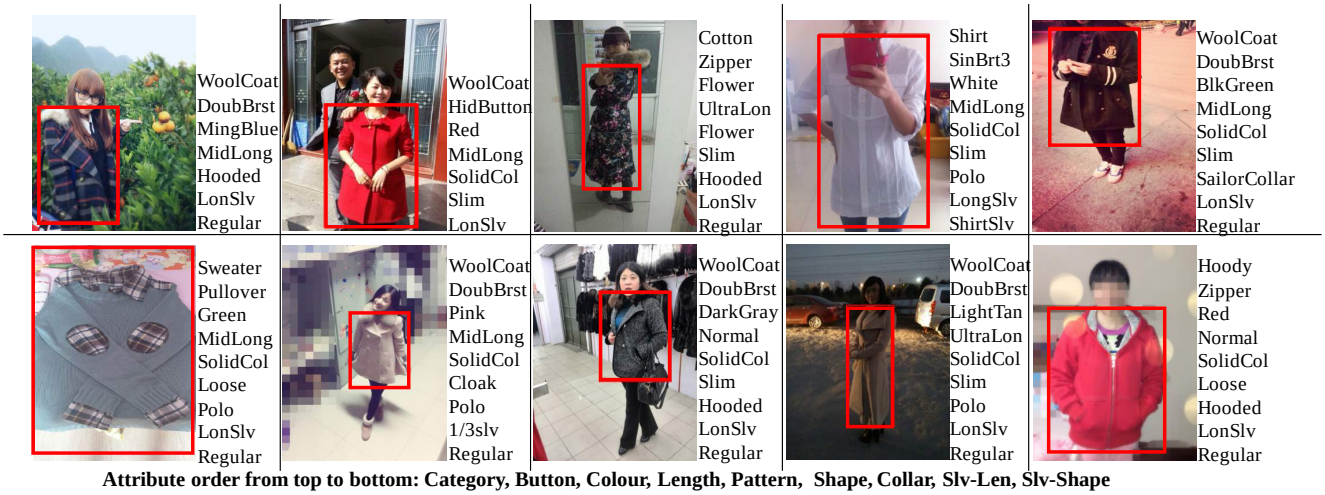
Figure 4. Failure cases: Incorrect attribute predictions (green) are shown against the corresponding ground truth (red) underneath.

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^{cls}. 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

Table 2. Evaluating the effects of multi-task and curriculum transfer learning in MTCT.

Method	Category	Button	Colour	Length	Pattern	Shape	Collar	Slv-Len	Slv-Shp	mAP^{cls}	mP^{ins}	mR^{ins}
JAN(NoAdpt) [27]	34.08	35.87	43.08	44.21	63.76	43.40	40.50	78.13	71.08	50.46	50.39	58.40
MTN(NoAdpt)	35.77	33.77	44.13	44.76	65.26	45.75	40.85	79.76	72.40	51.38	51.82	60.00
MTN(UD)	54.10	40.65	57.88	51.35	67.80	49.79	49.09	83.61	74.60	58.76	60.16	70.00
MTN(FTT)	61.92	42.65	65.43	55.16	70.06	49.00	50.55	85.54	76.04	61.82	62.53	72.76
MTCT	**65.96**	**43.57**	**66.86**	**58.27**	**70.55**	**51.40**	**58.97**	**86.05**	**77.54**	**64.35**	**64.97**	**75.66**



Attribute order from top to bottom: Category, Button, Colour, Length, Pattern, Shape, Collar, Slv-Len, Slv-Shape

Figure 3. A qualitative evaluation of our proposed MTCT method on unconstrained consumer images.

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 $36,769$ 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.

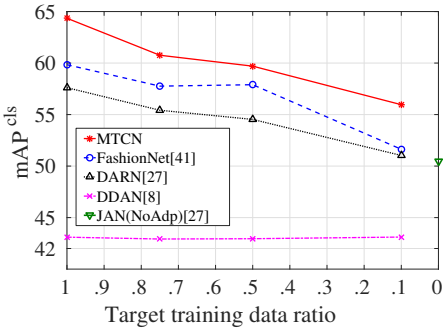


Figure 5. Model robustness vs. target training data size.

Table 3. Comparing curriculum vs. end-to-end transfer learning using the MTN network.

Method	Category	Button	Colour	Length	Pattern	Shape	Collar	Slv-Len	Slv-Shp	mAP^{cls}	mP^{ins}	mR^{ins}
End-to-End	61.11	41.39	63.66	56.29	70.02	51.39	55.45	84.69	76.69	62.30	63.00	73.37
Curriculum	**65.96**	**43.57**	**66.86**	**58.27**	**70.55**	**51.40**	**58.97**	**86.05**	**77.54**	**64.35**	**64.97**	**75.66**

Table 4. Comparing different loss functions in the MTCT network.

Method	Category	Button	Colour	Length	Pattern	Shape	Collar	Slv-Len	Slv-Shp	mAP^{cls}	mP^{ins}	mR^{ins}
triplet ranking [27]	63.57	42.01	63.80	56.16	69.37	50.58	57.03	85.24	75.60	62.60	63.45	73.83
t-STE [53]	**65.96**	**43.57**	**66.86**	**58.27**	**70.55**	**51.40**	**58.97**	**86.05**	**77.54**	**64.35**	**64.97**	**75.66**

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.

Figure 6. Examples of attribute based automatic clothing detection and matching in-the-wild (bottom) given clean shop/model samples (top), or vice versa. Each pair of images in each column is of the same clothing item matched from different domains likely on different people.

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 $\sim 8\%$ when only $<4,000$ target training images are available.

References

[1] R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *The Journal of Machine Learning Research*, 6(Nov):1817–1853, 2005.

[2] Y. Bengio, F. Bastien, A. Bergeron, N. Boulanger-Lewandowski, T. M. Breuel, Y. Chherawala, M. Cisse, M. Côté, D. Erhan, J. Eustache, et al. Deep learners benefit more from out-of-distribution examples. In *International Conference on Artificial Intelligence and Statistics*, pages 164–172, 2011.

[3] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[4] Y. Bengio et al. Deep learning of representations for unsupervised and transfer learning. *International Conference on Machine learning*, 27:17–36, 2012.

[5] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *International Conference on Machine learning*, pages 41–48, 2009.

[6] L. Bossard, M. Dantone, C. Leistner, C. Wengert, T. Quack, and L. Van Gool. Apparel classification with style. In *Asian Conference on Computer Vision*, pages 321–335. Springer, 2013.

[7] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *European Conference on Computer Vision*, pages 609–623. Springer, 2012.

[8] Q. Chen, J. Huang, R. Feris, L. M. Brown, J. Dong, and S. Yan. Deep domain adaptation for describing people based on fine-grained clothing attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5315–5324, 2015.

[9] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *IEEE International Conference on Computer Vision*, pages 1409–1416, 2013.

[10] Y. Deng, P. Luo, C. C. Loy, and X. Tang. Pedestrian attribute recognition at far distance. In *ACM International Conference on Multimedia*, pages 789–792, 2014.

[11] Z. Ding, N. M. Nasrabadi, and Y. Fu. Task-driven deep transfer learning for image classification. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2414–2418, 2016.

[12] S. K. Divvala, A. Farhadi, and C. Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3270–3277, 2014.

[13] J. L. Elman. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99, 1993.

[14] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[15] T. Evgeniou and M. Pontil. Regularized multi–task learning. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 109–117, 2004.

[16] R. Feris, R. Bobbitt, L. Brown, and S. Pankanti. Attribute-based people search: Lessons learnt from a practical surveillance system. In *ACM International Conference on Multimedia Retrieval*, page 153, 2014.

[17] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *IEEE International Conference on Computer Vision*, pages 2960–2967, 2013.

[18] J. Fu, J. Wang, Z. Li, M. Xu, and H. Lu. Efficient clothing retrieval with semantic-preserving visual phrases. In *Asian Conference on Computer Vision*, pages 420–431, 2012.

[19] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Transductive multi-view zero-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(11):2332–2345, 2015.

[20] A. C. Gallagher and T. Chen. Clothing cosegmentation for recognizing people. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[21] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

[22] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2066–2073, 2012.

[23] S. Gong, M. Cristani, S. Yan, and C. C. Loy. *Person re-identification*, volume 1. Springer, 2014.

[24] 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*, 36(11):2288–2302, 2014.

[25] J. Hoffman, E. Tzeng, J. Donahue, Y. Jia, K. Saenko, and T. Darrell. One-shot adaptation of supervised deep convolutional models. *arXiv e-prints*, 2013.

[26] J. Hu, J. Lu, and Y.-P. Tan. Deep transfer metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 325–333, 2015.

[27] J. Huang, R. S. Feris, Q. Chen, and S. Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *IEEE International Conference on Computer Vision*, 2015.

[28] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM International Conference on Multimedia*, pages 675–678, 2014.

[29] M. H. Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg. Where to buy it: Matching street clothing photos in online shops. In *IEEE International Conference on Computer Vision*, 2015.

[30] M. H. Kiapour, K. Yamaguchi, A. C. Berg, and T. L. Berg. Hipster wars: Discovering elements of fashion styles. In *European Conference on Computer Vision*, pages 472–488, 2014.

[31] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Image search with relative attribute feedback. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2973–2980, 2012.

[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[33] K. A. Krueger and P. Dayan. Flexible shaping: How learning in small steps helps. *Cognition*, 110(3):380–394, 2009.

[34] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *IEEE International Conference on Computer Vision*, pages 365–372, 2009.

[35] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958, 2009.

[36] R. Layne, T. M. Hospedales, S. Gong, and Q. Mary. Person re-identification by attributes. In *British Machine Vision Conference*, 2012.

[37] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan. Deep human parsing with active template regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(12):2402–2414, 2015.

[38] M. Lin, Q. Chen, and S. Yan. Network in network. *arXiv e-prints*, 2013.

[39] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3330–3337, 2012.

[40] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1096–1104, 2016.

[41] 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*, 24(12):5479–5491, 2015.

[42] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015.

[43] D. L. Rohde and D. C. Plaut. Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, 72(1):67–109, 1999.

[44] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele. What helps where–and why? semantic relatedness for knowledge transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 910–917, 2010.

[45] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[46] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *International Conference on Learning Representations*, 2014.

[47] L. Shao, F. Zhu, and X. Li. Transfer learning for visual categorization: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 26(5):1019–1034, 2015.

[48] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 806–813, 2014.

[49] B. Siddiquie, R. S. Feris, and L. S. Davis. Image ranking and retrieval based on multi-attribute queries. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 801–808, 2011.

[50] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[51] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1521–1528, 2011.

[52] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *IEEE International Conference on Computer Vision*, pages 4068–4076, 2015.

[53] L. Van Der Maaten and K. Weinberger. Stochastic triplet embedding. In *IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6, 2012.

[54] D. Vaquero, R. S. Feris, D. Tran, L. Brown, A. Hampapur, M. Turk, et al. Attribute-based people search in surveillance environments. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1–8, 2009.

[55] X. Wang and T. Zhang. Clothes search in consumer photos via color matching and attribute learning. In *ACM International Conference on Multimedia*, pages 1353–1356, 2011.

[56] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3570–3577, 2012.

[57] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, and J. Kim. Rotating your face using multi-task deep neural network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 676–684, 2015.

[58] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, pages 3320–3328, 2014.

[59] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*, pages 94–108, 2014.