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

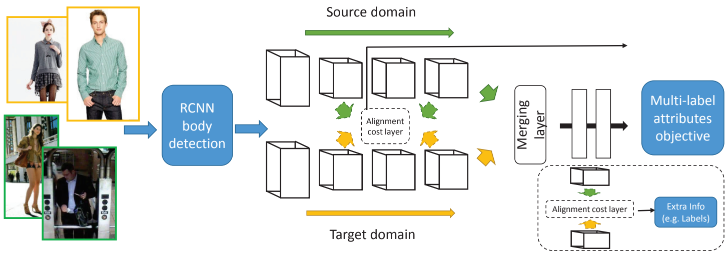

Figure 1: Our proposed Deep Domain Adaptation Network (DDAN). Source and target domains are modeled jointly with knowledge transfer occurring at multiple levels of the hierarchy through alignment cost layers.

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 (https://bankrobbers.fbi.gov/) 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").

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. [7] achieves impressive attribute prediction performance in unconstrained scenarios, but only considers nine human attributes. Existing systems for fashion analysis [1, 4, 6] and people search in surveillance videos [2, 5] also rely on a relatively small set of clothing attributes. Our work instead addresses the problem of describing people with very fine-grained clothing attributes. A natural question that arises in this setting is how to obtain a sufficient number of training samples for each attribute without significant annotation cost.

Data collection: We observe that online shopping stores such as Amazon.com and TMALL.com have a large set of garment images with associated descriptions. We created a huge dataset of clothing images with fine-grained attribute labels by crawling data from these shopping websites. Our dataset contains 1,108,013 clothing images with 25 different kinds attribute categories (e.g. type, color, pattern, season, occasion). The attribute labels are very fine-detailed. For instance, we can find thousands of different *values* for the "color" category. After data curation, we considered a subset of this data that is meaningful from our application perspective.

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.

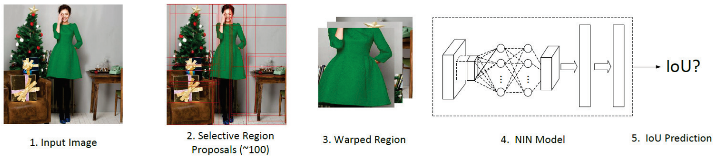

Figure 2: Enhanced R-CNN detection pipeline.

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. $f(X_s, X_t) = max(X_s, X_t)$, where X_s and X_t are the feature representations of the source and target connecting layers, respectively. So we can simply drop out this layer at testing time, when just the target domain images are available.

Our proposed deep domain adaptation network (DDAN) achieves better performance than traditional domain adaptation methods based on deep learning, such as backpropagation fine-tuning at a lower learning rate and re-training of the last network layer.

R-CNN body detector: Our body detection module is based on the R-CNN framework [3], with several enhancements made specifically for the clothing detection problem (see Figure 2). It consists of three submodules. 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. Our method achieves better results when compared to traditional R-CNN and deformable part-based models.

In summary, 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-world scenario.

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

[2] Rogerio Feris, Russel Bobbitt, Lisa Brown, and Sharath Pankanti. Attribute-based people search: Lessons learnt from a practical surveillance system. In *ICMR*, 2014.

[3] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv preprint arXiv:1311.2524*, 2013.

[4] S. Liu, L. Liu, and S. Yan. Fashion Analysis: Current Techniques and Future Directions. *IEEE Multimedia*, 2014.

[5] Daniel A Vaquero, Rogério Schmidt Feris, Duan Tran, Lisa Brown, Arun Hampapur, and Matthew Turk. Attribute-based people search in surveillance environments. In *WACV*, 2009.

[6] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg. Parsing clothing in fashion photographs. In *CVPR*, 2012.

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

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

* indicates equal contributions