DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations

Ziwei Liu[1] Ping Luo[3,1] Shi Qiu[2] Xiaogang Wang[1,3] Xiaoou Tang[1,3]
[1]The Chinese University of Hong Kong [2]SenseTime Group Limited [3]Shenzhen Institutes of Advanced Technology, CAS
{lz013,pluo,xtang}@ie.cuhk.edu.hk, sqiu@sensetime.com, xgwang@ee.cuhk.edu.hk

Abstract

Recent advances in clothes recognition have been driven by the construction of clothes datasets. Existing datasets are limited in the amount of annotations and are difficult to cope with the various challenges in real-world applications. In this work, we introduce DeepFashion[1], a large-scale clothes dataset with comprehensive annotations. It contains over 800,000 images, which are richly annotated with massive attributes, clothing landmarks, and correspondence of images taken under different scenarios including store, street snapshot, and consumer. Such rich annotations enable the development of powerful algorithms in clothes recognition and facilitating future researches. To demonstrate the advantages of DeepFashion, we propose a new deep model, namely FashionNet, which learns clothing features by jointly predicting clothing attributes and landmarks. The estimated landmarks are then employed to pool or gate the learned features. It is optimized in an iterative manner. Extensive experiments demonstrate the effectiveness of FashionNet and the usefulness of DeepFashion.

1. Introduction

Recently, extensive research efforts have been devoted to clothes classification [11, 1, 29], attribute prediction [3, 13, 4, 24], and clothing item retrieval [17, 6, 10, 27, 15], because of their potential values to the industry. However, clothes recognition algorithms are often confronted with three fundamental challenges when adopted in real-world applications [12]. First, clothes often have large variations in style, texture, and cutting, which confuse existing systems. Second, clothing items are frequently subject to deformation and occlusion. Third, clothing images often exhibit serious variations when they are taken under different scenarios, such as selfies *vs*. online shopping photos.

Previous studies tried to handle the above challenges by annotating clothes datasets either with semantic attributes (*e.g*. color, category, texture) [1, 3, 6], clothing locations (*e.g*. masks of clothes) [20, 12], or cross-domain image correspondences [10, 12]. However, different datasets are annotated with different information. A unified dataset with all the above annotations is desired. This work fills in this gap. As illustrated in Fig.1, we show that clothes recognition can benefit from learning these annotations jointly. In Fig.1 (a), given the additional landmark locations may improve recognition. As shown in Fig.1 (b), massive attributes lead to better partition of the clothing feature space, facilitating the recognition and retrieval of cross-domain clothes images.

Figure 1. (a) Additional landmark locations improve clothes recognition. (b) Massive attributes lead to better partition of the clothing feature space.

To facilitate future researches, we introduce *DeepFashion*, a comprehensively annotated clothes dataset that contains massive attributes, clothing landmarks, as well as cross-pose/cross-domain correspondences of clothing pairs. This dataset enjoys several distinct advantages over its precedents. (1) *Comprehensiveness* - images of DeepFashion are richly annotated with categories, attributes, land-

[1]The dataset is available at: http://mmlab.ie.cuhk.edu.hk/projects/DeepFashion.html

1063-6919/16 $31.00 © 2016 IEEE
DOI 10.1109/CVPR.2016.124

	DCSA [3]	ACWS [1]	WTBI [12]	DDAN [4]	DARN [10]	**DeepFashion**
# images	1856	145,718	78,958	341,021	182,780	**>800,000**
# categories + attributes	26	15	11	67	179	**1,050**
# exact pairs	N/A	N/A	39,479	N/A	91,390	**>300,000**
localization	N/A	N/A	bbox	N/A	N/A	**4∼8 landmarks**
public availability	yes	yes	no	no	no	**yes**

Table 1. Comparing DeepFashion with other existing datasets. DeepFashion offers the largest number of images and annotations.

marks, and cross-pose/cross-domain pair correspondences. It has 50 fine-grained categories and $1,000$ attributes, which are one order of magnitude larger than previous works [3, 4, 10]. Our landmark annotation is at a finer level than existing bounding-box label [12]. Such comprehensive and rich information are not available in existing datasets. (2) *Scale* - DeepFashion contains over $800K$ annotated clothing images, doubling the size of the largest one in the literature. 3) *Availability* - DeepFashion will be made public to the research community. We believe this dataset will greatly benefits the researches in clothes recognition and retrieval.

Meanwhile, DeepFashion also enables us to rigorously benchmark the performance of existing and future algorithms for clothes recognition. We create three benchmarks, namely clothing attribute prediction, in-shop clothes retrieval, and cross-domain clothes retrieval, *a.k.a.* street-to-shop. With such benchmarks, we are able to make direct comparisons between different algorithms and gain insights into their pros and cons, which we hope will eventually foster more powerful and robust clothes recognition and retrieval systems.

To demonstrate the usefulness of DeepFashion, we design a novel deep learning structure, FashionNet, which handles clothing deformation/occlusion by pooling/gating feature maps upon estimated landmark locations. When supervised by massive attribute labels, FashionNet learns more discriminative representations for clothes recognition. We conduct extensive experiments on the above benchmarks. From the experimental results with the proposed deep model and the state-of-the-arts, we show that the DeepFashion dataset promises more accurate and reliable algorithms in clothes recognition and retrieval.

This work has three main **contributions**. (1) We build a large-scale clothes dataset of over 800K images, namely DeepFashion, which is comprehensively annotated with categories, attributes, landmarks, and cross-pose/cross-domain pair correspondences. To our knowledge, it is the largest clothes dataset of its kind. (2) We develop FashionNet to jointly predict attributes and landmarks. The estimated landmarks are then employed to pool/gate the learned features. It is trained in an iterative manner. (3) We carefully define benchmark datasets and evaluation protocols for three widely accepted tasks in clothes recognition and retrieval. Through extensive experiments with our proposed model as well as other state-of-the-arts, we demonstrate the effectiveness of DeepFashion and FashionNet.

1.1. Related Work

Clothing Datasets As summarized in Table 1, existing clothes recognition datasets vary in size as well as the amount of annotations. The previous datasets were labeled with limited number of attributes, bounding boxes [12], or consumer-to-shop pair correspondences [10]. DeepFashion contains $800K$ images, which are annotated with 50 categories, $1,000$ attributes, clothing landmarks (each image has $4 \sim 8$ landmarks), and over $300K$ image pairs. It is the largest and most comprehensive clothes dataset to date. Some other datasets in the vision community were dedicated to the tasks of clothes segmentation, parsing [32, 31, 23, 16, 33] and fashion modeling [24, 30], while DeepFashion focuses on clothes recognition and retrieval.

Clothing Recognition and Retrieval Earlier works [28, 3, 7, 1, 6] on clothing recognition mostly relied on hand-crafted features, such as SIFT [19], HOG [5] and color histogram *etc*. The performance of these methods were limited by the expressive power of these features. In recent years, a number of deep models have been introduced to learn more discriminative representation in order to handle cross-scenario variations [10, 12]. Although these methods achieved good performance, they ignored the deformations and occlusions in the clothing images, which hinder further improvement of the recognition accuracy. FashionNet handles such difficulties by explicitly predicting clothing landmarks and pooling features over the estimated landmarks, resulting in more discriminative clothes representation.

2. The DeepFashion Dataset

We contribute DeepFashion, a large-scale clothes dataset, to the community. DeepFashion has several appealing properties. First, it is the largest clothing dataset to date, with over $800,000$ diverse fashion images ranging from well-posed shop images to unconstrained consumer photos, making it twice the size of the previous largest clothing dataset. Second, DeepFashion is annotated with rich information of clothing items. Each image in this dataset is labeled with 50 categories, $1,000$ descriptive attributes, and clothing landmarks. Third, it also contains over $300,000$ cross-pose/cross-domain image pairs. Some

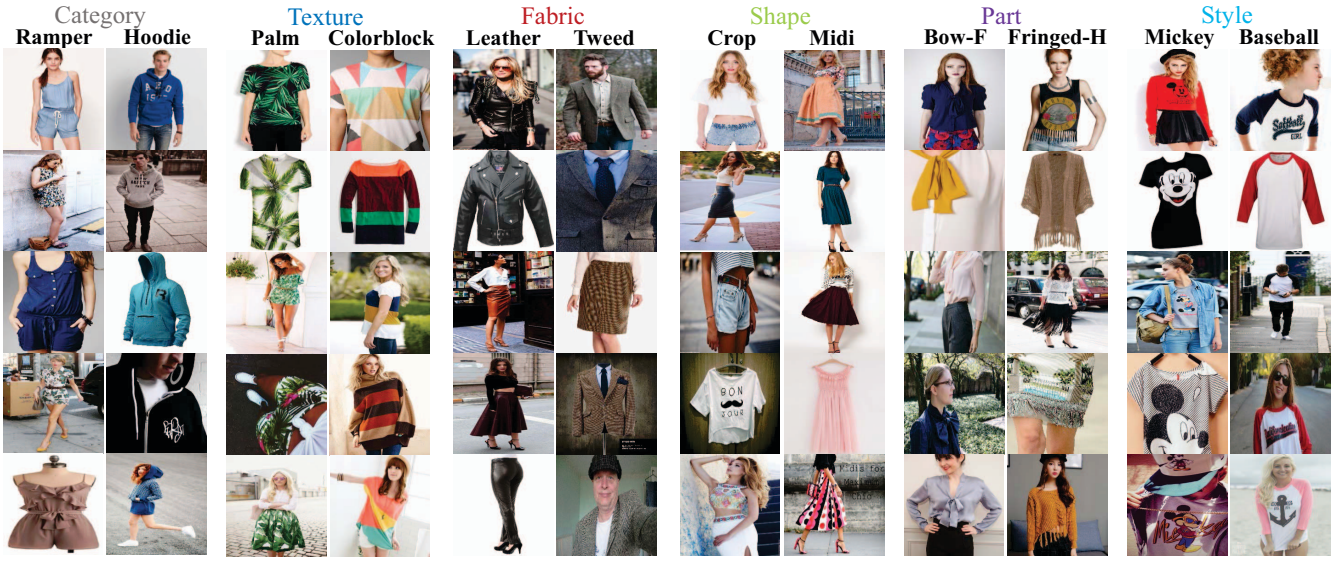

Figure 2. Example images of different categories and attributes in DeepFashion. The attributes form five groups: texture, fabric, shape, part, and style.

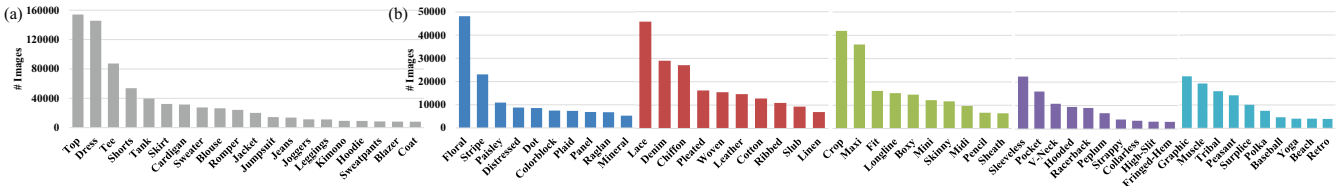

Figure 3. (a) Image number of the top-20 categories. (b) Image number of the top-10 attributes in each group.

example images along with the annotations are shown in Fig.2. From the comparison items summarized in Table 1, we see that DeepFashion surpasses the existing datasets in terms of scale, richness of annotations, as well as availability.

2.1. Image Collection

Shopping websites are a common source for constructing clothing datasets [10, 12]. In addition to this source, we also collect clothing images from image search engines, where the resulting images come from blogs, forums, and the other user-generated contents, which supplement and extend the image set collected from the shopping websites.

Collecting Images from Shopping Websites We crawled two representative online shopping websites, *Forever21*[2] and *Mogujie*[3]. The former one contains images taken by the online store. Each clothing item has $4 \sim 5$ images of varied poses and viewpoints. The latter one contains images taken by both the stores and consumers. Each clothing image in shop is accompanied by several user-taken photos of exactly the same clothing item. Therefore, these data not only cover the image distribution of professional online retailer stores, but also the other different domains such as street snapshots and selfies. We collected $1,320,078$ images of $391,482$ clothing items from these two websites.

Collecting Images from Google Images[4] To obtain meaningful query keywords for clothing images, we traversed the catalogue of several online retailer stores and collected names of clothing items, such as "animal print dress". This process results in a list of $12,654$ unique queries. We then feed this query set to Google Images, and download the returned images along with their associated meta-data. A total of $1,273,150$ images are collected from Google Images.

Data Cleaning We identified near- and exact-duplicate images by comparing fc7-responses after feeding them into AlexNet [14]. After the removal of the duplicates, we ask human annotators to remove unusable images that are of low resolution, image quality, or whose dominant objects are irrelevant to clothes. In total, $800,000$ clothing images are kept to construct DeepFashion.

2.2. Image Annotation

We advocate the following labeled information in order to aid the tasks for clothing recognition and retrieval. They are: (1) *Massive attributes* - this type of information is essential to recognize and represent the enormous clothing items; (2) *Landmarks* - the landmark locations can effectively deal with deformation and pose variation; (3) *Consumer-to-shop pairs* - these data is of great help in

[2] www.forever21.com
[3] www.mogujie.com
[4] https://www.google.com/imghp

bridging the cross-domain gap.

Generating Category and Attribute Lists We generated category and attribute lists from the query set collected in Sec.2.1, where most queries are of the form "adjective + noun" (*e.g.* "animal print dress"). For clothing categories, we first extracted the *nouns* (*e.g.* "dress") from the query set, resulting in 50 unique names of fine-grained categories. Next, we collected and merged the adjectives (*e.g.* "animal print"), and picked the top $1,000$ tags with highest frequency as the attributes. These attributes were categorized into five groups, characterizing texture, fabric, shape, part, and style, respectively.

Category and Attribute Annotation The category set is of moderate size (*i.e.* 50) and the category labels are mutually exclusive by definition. Therefore, we instruct professional human annotators to manually assign them to the images. Each image received at most one category label. The numbers of images for the top-20 categories are shown in Fig.3 (a). As for the $1,000$ attributes, since the number is huge and multiple attributes can fire on the same image, manual annotation is not manageable. We thus resort to the meta-data for automatically assigning attribute labels. Specifically, for each clothing image, we compare the attribute list with its associated meta-data, which is provided by Google or corresponding shopping website. We regard an attribute as "fired" if it is successfully matched in the image's meta-data. We show sample images for a number of selected attributes in Fig.2. We enumerated top ten attributes in each group, along with their image numbers in Fig.3 (b).

Landmark Annotation We define a set of clothing landmarks, which corresponds to a set of key-points on the structures of clothes. For instance, the landmarks for upper-body items are defined as left/right collar end, left/right sleeve end, and left/right hem. Similarly, we define landmarks for lower-body items and full-body items. As the definitions are straightforward and natural to average people, the human labelers could easily understand the task after studying a score of examples. As some of the landmarks are frequently occluded in images, we also labeled the visibility (*i.e.* whether a landmark is occluded or not) of each landmark. Note that our landmarks are clothes-centric, and thus different from joints of human body. Fig.4 illustrates some examples of landmark annotations.

Pair Annotation As discussed in Sec.2.1, we collected clothing images under different domains, including photos from web stores, street snapshots, and consumers. We clean such image pairs by removing noisy images, ensuring the accuracy of our pairwise correspondences.

Quality Control We took the following steps to control the labeling quality. (1) We discarded images with too few textual meta-data. (2) After automatically annotating attributes, human annotators also conducted a fast screening

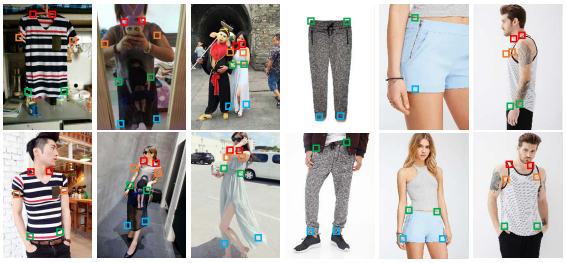

Figure 4. Landmarks and pair annotation in DeepFashion. Landmarks are defined for upper-body clothes, lower-body clothes and full-body clothes, respectively. Images in the same column capture the same clothing item.

to rule out falsely "fired" images for each attribute to ensure the precision. (3) For other manually annotated labels, we collected annotations from two different annotators and check their consistency. Around 0.5% samples were found inconsistent and required further labelling from a third annotator.

2.3. Benchmarks

We build the following benchmarks out of DeepFashion for evaluating different methods.

Category and Attribute Prediction This task is to classify 50 fine-grained categories and $1,000$ attributes. There are $63,720$ diverse images in this benchmark. For category classification, we employ the standard top-k classification accuracy as evaluation metric. For attribute prediction, our measuring criteria is the top-k recall rate following [9], which is obtained by ranking the $1,000$ classification scores and determine how many attributes have been matched in the top-k list.

In-Shop Clothes Retrieval This task is to determine if two images taken in shop belong to the same clothing item [5]. It is important when customers encounter shop image on photo sharing sites and would like to know more about its item information on online retailer stores. This benchmark contains $54,642$ images of $11,735$ clothing items from *Forever21*. Top-k retrieval accuracy is adopted to measure the performance of fashion retrieval, such that a successful retrieval is counted if the exact fashion item has been found in the top-k retrieved results.

Consumer-to-Shop Clothes Retrieval This scenario has been considered by several previous works [10, 12], aiming at matching consumer-taken photos with their shop counterparts. We select $251,361$ consumer-to-shop image pairs from *Mogujie* for this benchmark. Again, top-k retrieval accuracy is employed to evaluate performance.

3. Our Approach

To demonstrate the usefulness of DeepFashion, we propose a novel deep model, FashionNet, which simultane-

[5]We further annotate each image with its scale (zoom-in/zoom-out) and pose (front-view/side-view) using meta-data, which can be used for analyzing the influence of different clothing variations.

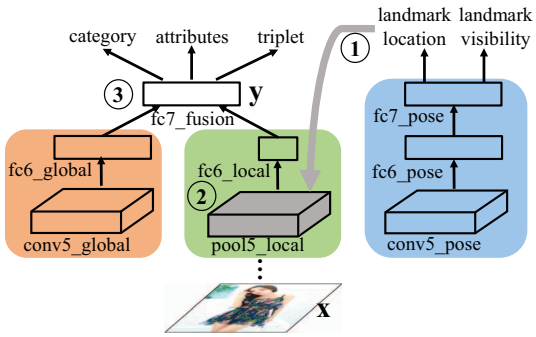

Figure 5. Pipeline of FashionNet, which consists of global appearance branch (in orange), local appearance branch (in green) and pose branch (in blue). Shared convolution layers are omitted for clarity.

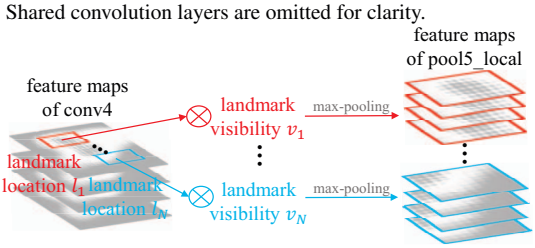

Figure 6. Schematic illustration of landmark pooling layer.

ously predicts landmarks and attributes. The estimated landmark locations are then employed to pool/gate the learned features, inducing discriminative clothing features. This procedure performs in an iterative manner. And the whole system can be learned end-to-end.

Network Structure The network structure of FashionNet is similar to VGG-16 [25], which has been demonstrated powerful in various vision tasks such as object recognition [25] and segmentation [18]. Specifically, the structures of FashionNet below the penultimate (*i.e.* from top to bottom) convolutional layer are the same as VGG-16, except the last convolutional layer, which is carefully design for clothes. As illustrated in Fig.5, the last convolutional layer in VGG-16 is replaced by three branches of layers, highlighted in red, green, and blue respectively. The branch in red captures global features of the entire clothing item, while the branch in green captures local features pooling over the estimated clothing landmarks. The branch in blue predicts the landmarks' locations as well as their visibility (*i.e.* whether they have been occluded or not). Moreover, the outputs of the branches in red and green are concatenated together as in "fc7_fusion" to jointly predict the clothes categories, attributes, and to model clothes pairs.

Forward Pass The forward pass of FashionNet consists of three stages as shown in Fig.5. At the first stage, a clothes image is fed into the network and passed through the branch in blue, so as to predict the landmarks' locations. At the second stage, the estimated landmarks are employed to pool or gate the features in "pool5_local", which is a *landmark pooling layer*, leading to local features that are invariant to deformations and occlusions of clothes. At the third stage, the global features of "fc6_global" and the landmark-pooled local features of "fc6_local" are concatenated together in "fc7_fusion".

Backward Pass The backward pass back-propagates the errors of four kinds of loss functions in an iterative manner. Here, we first introduce these loss functions and then discuss the iterative training strategy. These loss functions include a regression loss for landmark localization, a softmax loss for the predictions of landmark visibility and clothes categories, a cross-entropy loss for attribute predictions, and finally a triplet loss for metric learning of the pairwise clothes images. First, a modified L_2 regression loss is used to localize landmarks, $L_{landmarks} = \sum_{j=1}^{|D|} \|\mathbf{v}_j \cdot (\hat{\ell}_j - \ell_j)\|_2^2$, where D, $\hat{\ell}_j$, and $\mathbf{v}_j$ denote the number of training samples, the ground truth locations of the landmarks of the j-th sample, and a vector of its landmarks' visibility, respectively. Unlike the conventional regression loss, the visibility variables remedy missing ground truth locations of the landmarks, in the sense that the error is not propagated back when a landmark is occluded. Second, we adopt 1-of-K softmax loss to classify landmark visibility and fine-grained categories, denoted as $L_{visibility}$ and $L_{category}$ respectively.

Third, a weighted cross-entropy loss is utilized to predict attributes

$$L_{attributes} = \sum_{j=1}^{|D|} \big(w_{pos} \cdot \mathbf{a}_j \log p(\mathbf{a}_j|\mathbf{x}_j) + w_{neg} \cdot (1 - \mathbf{a}_j) \log(1 - p(\mathbf{a}_j|\mathbf{x}_j))\big), \quad (1)$$

where $\mathbf{x}_j$ and $\mathbf{a}_j$ represent the j-th clothes image and its attribute labels. w_{pos} and w_{neg} are two coefficients, determined by the ratio of the numbers of positive and negative samples in the training set.

Fourth, to learn the metric described by clothes pairs, we employ triplet loss introduced in [22], which enforces distance constraints among positive and negative samples

$$L_{triplet} = \sum_{j=1}^{|D|} \max\{0, m + d(\mathbf{x}_j, \mathbf{x}_j^+) - d(\mathbf{x}_j, \mathbf{x}_j^-)\}, \quad (2)$$

where a three-tuple $(\mathbf{x}, \mathbf{x}^+, \mathbf{x}^-)$ is a triplet. $\mathbf{x}^+$ and $\mathbf{x}^-$ indicate clothes images of the same and different item with respect to $\mathbf{x}$. $d(\cdot, \cdot)$ is a distance function and m is the margin parameter.

FashionNet is optimized by weighted combing the above loss functions. Here we discuss the iterative training strategy that repeats the following two steps. *In the first step*, we treat the branch in blue as the main task and the remaining branches as the auxiliary tasks. To this end, we assign $L_{visibility}$ and $L_{landmark}$ with large weights, while the other loss functions have small weights. This is to train landmark estimation with the assistance of the other tasks, since they are correlated. Joint optimization leads to

	Category		Texture		Fabric		Shape		Part		Style		All	
	top-3	top-5	top-3	top-5	top-3	top-5	top-3	top-5	top-3	top-5	top-3	top-5	top-3	top-5
WTBI [3]	43.73	66.26	24.21	32.65	25.38	36.06	23.39	31.26	26.31	33.24	49.85	58.68	27.46	35.37
DARN [10]	59.48	79.58	36.15	48.15	36.64	48.52	35.89	46.93	39.17	50.14	66.11	71.36	42.35	51.95
FashionNet+100	47.38	70.57	28.05	36.59	29.12	40.58	28.51	36.51	31.65	38.53	53.92	62.47	31.58	39.06
FashionNet+500	57.44	77.39	34.73	46.35	34.47	46.60	33.61	44.57	38.48	49.01	63.48	67.94	38.94	49.71
FashionNet+Joints [34]	72.30	81.52	35.92	48.73	38.21	49.04	37.59	47.73	40.21	51.81	64.91	73.14	43.14	52.33
FashionNet+Poselets [34]	75.34	84.87	36.85	49.11	38.88	49.48	38.19	47.09	41.60	52.85	64.84	73.03	43.57	52.65
FashionNet (Ours)	**82.58**	**90.17**	**37.46**	**49.52**	**39.30**	**49.84**	**39.47**	**48.59**	**44.13**	**54.02**	**66.43**	**73.16**	**45.52**	**54.61**

Table 2. Performance of category classification and attribute prediction.

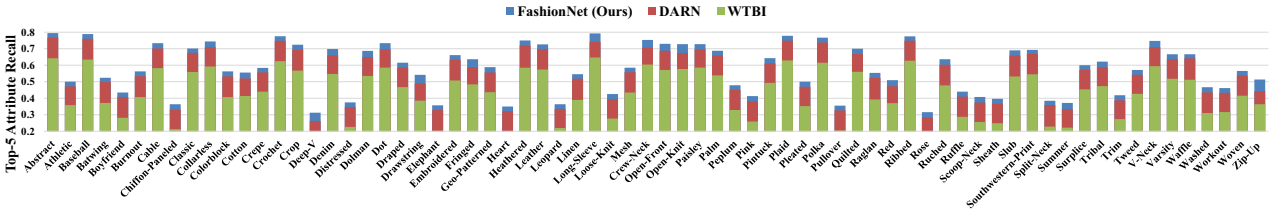

Figure 7. Per-attribute prediction performance for 70 representative attributes. FashionNet consistently outperforms WTBI [12] and DARN [10] on all attributes.

better convergence, which is demonstrated in Sec.4.2. *In the second step*, we predict clothing categories and attributes, as well as to learn the pairwise relations between clothes images. In this step, the estimated landmark locations are used to pool the local features. The above two steps are iterated until convergence. This procedure is similar to [21].

Landmark Pooling Layer The landmark pooling layer is a key component in FashionNet. Here, we discuss it in detail. As shown in Fig.6, the inputs of the landmark pooling layer are the feature maps (*i.e.* "conv4") and the estimated landmarks. For each landmark location ℓ, we first determine its visibility v. The responses of invisible landmark are gated to zero. Next, we perform max-pooling inside the region around ℓ to obtain local feature maps. These local feature maps are stacked to form the final feature maps of "pool5_local". The back-propagation of the landmark pooling layer is similar to the RoI pooling layer introduced in [8]. However, unlike [8] that treated the pooled regions independently, the landmark pooling layer captures interaction between clothing landmarks by concatenating local features.

4. Experiments

Data We pre-train FashionNet on a subset of $300,000$ images of DeepFashion, another subset of $50,000$ images is used as validation data. In testing, we employ part of the benchmark data to fine-tune the pre-trained models on the three benchmarks. We ensure that no fashion item overlaps between fine-tuning and testing sets.

Competing Methods We compare FashionNet with two recent deep models that showed compelling performance in clothes recognition, including Where To Buy It (WTBI) [12] and Dual Attribute-aware Ranking Network (DARN) [10]. Both of them are trained using clothes bounding boxes. Specifically, WTBI concatenated multi-layer perceptron (MLP) on top of the pre-trained ImageNet models [25]. We only implement the category-independent metric network of WTBI, which handles all clothing categories in a single network. DARN adopted an attribute-regularized two-stream CNN. One stream handles shop images, while the other handles street images. Note that for category classification and attribute prediction, only one stream of DARN is used. We train WTBI and DARN using the same amount of data and protocol as FashionNet did.

We also vary building blocks of FashionNet for an ablation study, including FashionNet+100 and FashionNet+500. They represent that we only utilize 100 and 500 attributes to learn FashionNet respectively, instead of 1,000 attributes used in the full model. Next, we replace fashion landmarks in our model with detected human joints [34] and poselets [2] to pool/gate features in the stages of training and test. They are denoted as FashionNet+Joints and FashionNet+Poselets, respectively.

4.1. Results

This section provides quantitative evaluations of different methods on the three benchmarks. We also investigate multiple building blocks of the proposed FashionNet. Table 2 summarizes the performance of different methods on category classification and attribute prediction.

Category Classification In fine-grained category classification, we have three observations. *First*, FashionNet significantly outperforms WTBI and DARN by 20 percent when evaluated using the top-3 accuracy. It outperforms them by 10 percent in the top-5 accuracy. Please refer to Sec.2.3 for the details of the evaluation metrics of the benchmark. These results show that by adding informative landmarks, FashionNet can better discover fine-grained clothing traits than existing deep models. *Second*, when replacing the clothing landmarks in FashionNet with human joints and poselets, 6∼9 percent performance drops are

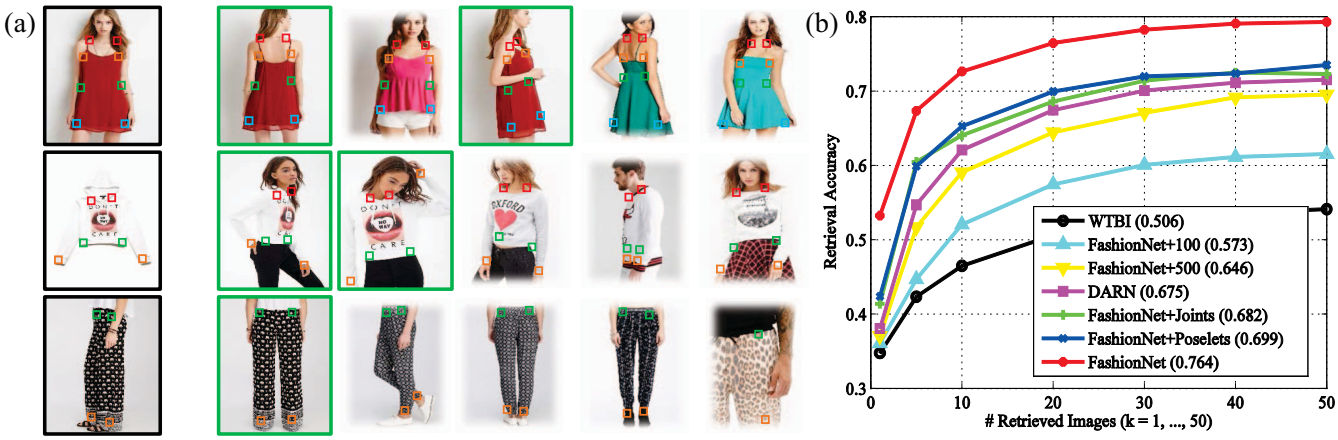

Figure 8. Results on in-shop clothes retrieval benchmark. (a) Example queries, top-5 retrieved images, along with their predicted landmarks. Correct matches are marked in green. (b) Retrieval accuracies of different methods under comparison.

observed. As the clothing landmarks are defined based on domain-specific semantics of clothes, they are more effective than human joints/poselets in clothes recognition. *Third*, using massive attributes benefits fine-grained category classification. When training FashionNet with different number of attributes, including 100, 500, and 1,000, the classification accuracies significantly increase. The full model surpasses FashionNet-500 and -100 by 13 and 20 percent in the top-5 accuracy respectively, showing that richer attribute information helps comprehensive profiling of different clothing variations.

Attribute Prediction For attribute prediction, similar conclusions can be drawn for the effectiveness of informative landmarks and massive attributes. To understand the strength of FashionNet, we also present the attribute recall results for each of the five attribute groups. We demonstrate that FashionNet achieves compelling results in the attribute groups of "shape" and "part", because discriminative information of these attributes normally exist around clothing landmarks, thus can be well captured by landmark pooling in FashionNet. Fig.7 illustrates the per-attribute recall rates of top-5 accuracy for the 70 representative attributes. Our approach consistently outperforms WTBI and DARN on all of the attributes in this benchmark.

In-Shop Clothes Retrieval Fig.8 shows the top-k retrieval accuracy of all the compared methods with k ranging from 1 to 50. We also list the top-20 retrieval accuracy after the name of each method. We can clearly see that our model (FashionNet) achieves best performance (0.764) among all the methods under comparison, while WTBI has the lowest accuracy (0.506). The poor performance of WTBI is as expected, since it directly used the pre-trained ImageNet features, which are not suitable to describe clothes. Notably, compared with DARN, FashionNet boots the top-20 accuracy from 0.675 to 0.764, and a 15 percent relative improvement is attained. This reveals the merits of employing landmarks to pool and gate learned features. When we replace clothing landmarks with human joints (FashionNet+Joints) or poselets (FashionNet+Poselets), the accuracy drops by 8 and 6 percent respectively, indicating such options are suboptimal. Compared with FashionNet+100 and FashionNet+500, FashionNet increase the accuracy by 19 and 12 percent, respectively, which highlights the effectiveness of using massive clothing attributes for training deep models. Some of the sample results are given in Fig.8 (a), where top retrieved images along with predicted landmark locations are shown.

Consumer-to-Shop Clothes Retrieval We show the detailed retrieval accuracy of different methods in Fig.9 (b). Compared with in-shop retrieval, methods on this benchmark achieve much lower accuracies, which reflect the inherent difficulty of consumer-to-shop clothes retrieval. Similar to in-shop clothes retrieval, FashionNet achieves the best top-20 accuracy (*i.e.* 0.188) among different methods. The relative improvement of FashionNet over DARN rises to 70 percent, compared to 15 percent of the previous benchmark, indicating the landmark locations are of greater importance for more complex scenarios. Besides, the retrieval accuracy increases when more training attributes are explored. Moreover, using human landmarks rather than clothing landmarks degrades the accuracy. These observations are consistent with those in the previous task. Some sample queries along with their top matches are shown in Fig.9 (a).

4.2. Further Analysis

As landmark estimation plays a key role in FashionNet, we conducted a quantitative evaluation of this component in order to better understand our method. For a more detailed analysis of search results, we also explore how different variations of clothes affect the retrieval accuracy.

Landmark Estimation Fig.10 (a) illustrates the detection rates over varying thresholding distances for different clothing landmarks. Similar to [26], percentage of detected

Figure 9. Results on consumer-to-shop clothes retrieval benchmark. (a) Example queries, top-3 retrieved images, along with their predicted landmarks. Correct matches are marked in green. (b) Retrieval accuracies of different methods under comparison.

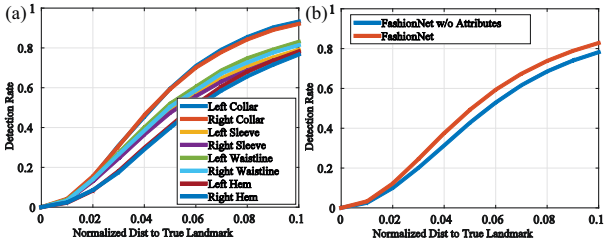

Figure 10. (a) Detection rates of different clothing landmarks. (b) Detection rates with and without using attributes.

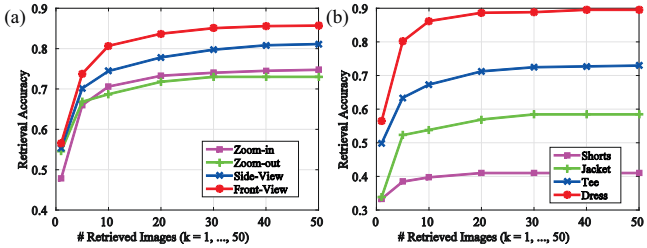

Figure 11. (a) Retrieval accuracies under different poses and scales. (b) Retrieval accuracies of different clothing categories.

joints (PDJ) is utilized to evaluate landmark estimation. When the normalized distance equals 0.1, the detection rates are above 80 percent for all the eight landmarks. We can further observe that detection rate of collars are higher than that of sleeves, waistlines, and hems. This is because collars are relatively rigid w.r.t. human's neck, whereas sleeves, waistlines, and hems are more flexible beyond common human joints. Fig.10 (b) demonstrates that rich attribute information facilitates landmark localization, because some attributes can effectively describe the appearance of certain clothing landmarks, such as "cap-sleeve" and "fringed-hem".

Variations of Clothes We choose the in-shop clothes retrieval benchmark to investigate the influence of different variations of clothes. Fig.11 (a) illustrates the retrieval performance of query images with different variations. We can see that scale variations are more challenging than pose variations. Another interesting observation is that zoom-in images perform worse than zoom-out images when $k = 1$, however its performance increases when k gets larger. It is because landmarks are essential for accurate fashion retrieval, but they are undetectable in zoom-in images. The fine-grained texture attributes help recognize zoom-in images and may guarantee an acceptable retrieval performance when k gets large. From Fig.11 (b), we can observe that "dress" has the highest accuracy while "shorts" has the lowest, because "dresses" generally have much more distinguishable features, such as local traits and colors. "Shorts", on the other hand, tend to have similar shape and relatively plain textures.

5. Conclusions

This work presents DeepFashion, a large-scale clothing dataset with comprehensive annotations. DeepFashion contains over 800,000 images, which are richly labeled with fine-grained categories, massive attributes, landmarks, and cross-pose/cross-domain image correspondence. It surpasses existing clothing datasets in terms of scale as well as richness of annotation. To demonstrate the advantages of such comprehensive annotations, we designed a novel deep model, namely FashionNet, that learns clothing features by jointly predicting landmark locations and massive attributes. The estimated landmarks are used to pool or gate the learned feature maps, which leads to robust and discriminative representations for clothes. We establish benchmark datasets for three widely accepted tasks in clothing recognition and retrieval. Through extensive experiments, we demonstrate the effectiveness of FashionNet and the usefulness of DeepFashion, which may significantly facilitate future researches.

Acknowledgement This work is partially supported by SenseTime Group Limited, the Hong Kong Innovation and Technology Support Programme (No. ITS/121/15FX), the General Research Fund sponsored by the Research Grants Council of Hong Kong (Project Nos. CUHK14203015, CUHK14207814), and the National Natural Science Foundation of China (61503366).

References

[1] L. Bossard, M. Dantone, C. Leistner, C. Wengert, T. Quack, and L. Van Gool. Apparel classification with style. In *ACCV*, pages 321–335. 2012. 1, 2

[2] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, pages 1365–1372, 2009. 6

[3] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *ECCV*, pages 609–623. 2012. 1, 2, 6

[4] 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 *CVPR*, pages 5315–5324, 2015. 1, 2

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005. 2

[6] W. Di, C. Wah, A. Bhardwaj, R. Piramuthu, and N. Sundaresan. Style finder: Fine-grained clothing style detection and retrieval. In *CVPR Workshops*, pages 8–13, 2013. 1, 2

[7] J. Fu, J. Wang, Z. Li, M. Xu, and H. Lu. Efficient clothing retrieval with semantic-preserving visual phrases. In *ACCV*, pages 420–431. 2012. 2

[8] R. Girshick. Fast r-cnn. In *ICCV*, 2015. 6

[9] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe. Deep convolutional ranking for multilabel image annotation. *arXiv preprint arXiv:1312.4894*, 2013. 4

[10] J. Huang, R. S. Feris, Q. Chen, and S. Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *ICCV*, 2015. 1, 2, 3, 4, 6

[11] Y. Kalantidis, L. Kennedy, and L.-J. Li. Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos. In *ICMR*, pages 105–112, 2013. 1

[12] 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 *ICCV*, 2015. 1, 2, 3, 4, 6

[13] M. H. Kiapour, K. Yamaguchi, A. C. Berg, and T. L. Berg. Hipster wars: Discovering elements of fashion styles. In *ECCV*, pages 472–488. 2014. 1

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012. 3

[15] X. Liang, L. Lin, W. Yang, P. Luo, J. Huang, and S. Yan. Clothes co-parsing via joint image segmentation and labeling with application to clothing retrieval. In *IEEE Transactions on Multimedia*, 2016. 1

[16] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, and S. Yan. Fashion parsing with weak color-category labels. *TMM*, 16(1):253–265, 2014. 2

[17] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *CVPR*, pages 3330–3337, 2012. 1

[18] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang. Semantic image segmentation via deep parsing network. In *ICCV*, 2015. 5

[19] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 2

[20] P. Luo, X. Wang, and X. Tang. Pedestrian parsing via deep decompositional network. In *ICCV*, pages 2648–2655, 2013. 1

[21] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 6

[22] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 5

[23] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun. A high performance crf model for clothes parsing. In *ACCV*, 2014. 2

[24] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun. Neuroaesthetics in fashion: Modeling the perception of beauty. In *CVPR*, 2015. 1, 2

[25] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5, 6

[26] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, pages 1653–1660, 2014. 7

[27] A. Veit, B. Kovacs, S. Bell, J. McAuley, K. Bala, and S. Belongie. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *ICCV*, pages 4642–4650, 2015. 1

[28] X. Wang and T. Zhang. Clothes search in consumer photos via color matching and attribute learning. In *ACM MM*, pages 1353–1356, 2011. 2

[29] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang. Learning from massive noisy labeled data for image classification. In *CVPR*, pages 2691–2699, 2015. 1

[30] K. Yamaguchi, T. L. Berg, and L. E. Ortiz. Chic or social: Visual popularity analysis in online fashion networks. In *ACM MM*, pages 773–776, 2014. 2

[31] K. Yamaguchi, M. H. Kiapour, and T. Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *ICCV*, pages 3519–3526, 2013. 2

[32] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *CVPR*, pages 3570–3577, 2012. 2

[33] W. Yang, P. Luo, and L. Lin. Clothing co-parsing by joint image segmentation and labeling. In *CVPR*, pages 3182–3189, 2014. 2

[34] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, pages 1385–1392, 2011. 6