Attentive Fashion Grammar Network for Fashion Landmark Detection and Clothing Category Classification

Wenguan Wang*[1,2], Yuanlu Xu*[2], Jianbing Shen†[1], and Song-Chun Zhu[2]

[1]Beijing Lab of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, China
[2]Department of Computer Science and Statistics, University of California, Los Angeles, USA

Abstract

This paper proposes a knowledge-guided fashion network to solve the problem of visual fashion analysis, e.g., fashion landmark localization and clothing category classification. The suggested fashion model is leveraged with high-level human knowledge in this domain. We propose two important fashion grammars: (i) dependency grammar capturing kinematics-like relation, and (ii) symmetry grammar accounting for the bilateral symmetry of clothes. We introduce Bidirectional Convolutional Recurrent Neural Networks (BCRNNs) for efficiently approaching message passing over grammar topologies, and producing regularized landmark layouts. For enhancing clothing category classification, our fashion network is encoded with two novel attention mechanisms, i.e., landmark-aware attention and category-driven attention. The former enforces our network to focus on the functional parts of clothes, and learns domain-knowledge centered representations, leading to a supervised attention mechanism. The latter is goal-driven, which directly enhances task-related features and can be learned in an implicit, top-down manner. Experimental results on large-scale fashion datasets demonstrate the superior performance of our fashion grammar network.

1. Introduction

With the rapid development of electronic commerce and the boom of online shopping, visual clothing analysis has attracted lots of interests in computer vision. More recently, benefited from the availability of large-scale fashion datasets, deep learning based models gained astonishing success in this area, such as clothing item retrieval [13, 19], and fashion image classification [38, 25, 28], to name a few.

In this paper, we address two key problems in visual fashion analysis, namely fashion landmark localization and clothing category classification. The success of previous deep learning based fashion models [19, 25, 22, 9] has proven the potential of applying neural network in this area. However, few of them attacked how to inject high-level human knowledge (such as geometric relationships among landmarks) into fashion models. In this paper, we propose a fashion grammar model that combines the learning power of neural network and domain-specific grammars that capture the kinematic and symmetric relations between clothing landmarks. For modeling the massage passing process over fashion grammars, we introduce a novel network architecture, Bidirectional Convolutional Recurrent Neural Network (BCRNN), which is flexible to our tree-structured models and generates more reasonable landmark layouts with global grammar constraints. Crucially, our whole deep grammar model is fully differentiable and can be trained in end-to-end manner.

This work also proposes two important attention mechanisms for boosting fashion image classification. The first one is *fashion landmark-aware*, which leverages the strong representation ability of fashion landmarks and can be learnt in supervised manner. This attention is able to generate landmark-aligned clothing features, which makes our model look for the informative semantic parts of garments. The second attention is *clothing category-driven* and trained in goal-driven way. Such attention mechanism learns to directly enhance task-related features and thus improves the classification performance. The attentions provide the model with more robust clothing representations and filter out useless information.

Comprehensive evaluations on two large-scale datasets

*Wenguan Wang and Yuanlu Xu contributed equally.

†Corresponding author: *Jianbing Shen* (shenjianbing@bit.edu.cn).
This work was supported by the Beijing Natural Science Foundation under Grant 4182056, the Fok Ying Tung Education Foundation under Grant 141067, and the Specialized Fund for Joint Building Program of Beijing Municipal Education Commission. This work was also supported by ONR MURI project N00014-16-1-2007, DARPA XAI Award N66001-17-2-4029, and NSF IIS 1423305.

[25, 26] demonstrate that our fashion grammar model outperforms the state-of-the-arts. Additionally, we experimentally demonstrate that our BCRNN based fashion grammars and attention modules give non-trivial improvements.

Contribution. Our main contribution is three-fold: i) We develop a deep grammar network to encode a set of knowledge over fashion clothes. The fashion knowledge, represented in grammar format, explicitly expresses the relations (*i.e.*, kinematics, and symmetry) of fashion landmarks, and serve as basis for constructing our fashion landmark detection module. ii) We present Bidirectional Convolutional Recurrent Neural Network (BCRNN) for approaching message passing over the suggested fashion grammars. The chain-structure topology of BCRNNs efficiently represents the rich relations of clothes, and our fashion model is fully differentiable which can be trained in end-to-end manner. iii) We introduce two attention mechanisms, one is landmark-aware and domain-knowledge-involved, and the other one directly focuses on the category relevant image regions and can be learned in goal driven manner.

2. Related Work

Visual fashion understanding has drawn lots of interests recently, due to its wide spectrum of human-centric applications such as clothing recognition [4, 25, 18, 15, 1], retrieval [44, 13, 23, 50], recommendation [21, 38, 25, 16], parsing [51, 53] and fashion landmark detection [25, 26]. *Earlier fashion models* [4, 21, 44, 23] are mostly relied on handcrafted features (*e.g.*, SIFT, HOG) and seek for powerful clothing representations, such as graph models [5], contextual information [40, 13], general object proposals [13], human parts [40, 23], bounding boxes [7] and semantic masks [50, 51, 53, 22, 12].

With the availability of large-scale fashion datasets [38, 25, 26], *deep learning based models* [19, 25, 26, 22, 28, 9] were proposed and outperformed prior work by a large margin. In particular, Huang *et al*. [19] introduced a Dual Attribute-aware Ranking Network (DARN) for clothing image retrieval. Liu *et al.* [25] proposed a branched neural network, for simultaneously performing clothing retrieval, classification, and landmark detection. More recently, in [26], a deep learning based model was designed as a combination of three cascaded networks for gradually refining fashion landmark estimates. Yan *et al*. [52] combined selective dilated convolution and recurrent spatial transformer for localizing cloth landmarks in unconstrained scenes. The success of those deep learning based fashion models demonstrate the strong representation power of neural network. However, they seldomly explore the rich domain-specific knowledge of clothes. In comparison, we propose a deep fashion grammar network that incorporates both powerful learning capabilities of neural networks and high-level semantic relations in visual fashion.

Grammar models in computer vision are powerful tool for modeling high-level human knowledge in specific domains, such as the decompositions of scenes [14, 24, 34], semantic relations between human and objects [56, 33], dependencies between human parts [47, 10], and the compatibility relations between human attributes over human hierarchy [48, 49, 31]. They are a natural choice for modeling rich relations and diverse structures in this world. Grammar models allow an expert inject domain-specific knowledge into the algorithms, thus avoiding local ambiguities and hard decisions [2, 31]. In this paper, we first propose two fashion grammars that account for dependent and symmetric relations in clothes. In particular, we ground these knowledge in a BCRNN based deep learning model which can be end-to-end trained with back-propagation.

Attention mechanism in computer vision has been popular in the tasks of image caption [46], Visual Question Answering (VQA) [37, 55], object detection [3, 45] and image recognition [43, 6, 42, 20]. Those methods show that top-down attention mechanism is effective as it allows the network to learn which regions in an image to attend to solve their tasks. In this paper, two kinds of attentions, namely category-directed and landmark-aware attentions, are proposed. As far as we know, no attention mechanism has been applied to feedforward network structure to achieve state-of-the-art results in visual fashion understanding tasks. Besides, in contrast to previous part-based fashion models [40, 23, 25, 26] with hard deterministic constraints in feature selection, our attentions act as soft constraints and can be leant in a stochastic way from data.

3. Our Approach

We first describe our fashion grammar network for fashion landmark detection (§3.1). Then we introduce two attention mechanisms for clothing image classification (§3.2).

3.1. Fashion Grammar Network for Fashion Landmark Detection

Problem Definition. Clothing landmark detection aims to predict the positions of K functional key points defined on the fashion items, such as the corners of neckline, hemline, and cuff. Given an image I, the goal is to predict cloth landmark locations L:

$$L = \{L_k : k = 1, \dots, K\},\ L_k \in \mathbb{R}^2, \tag{1}$$

where L_k can be any pixel locations (u, v) in an image.

Previous fashion landmark methods [25, 26] formulate this problem as regression. They train a deep learning model and use a function $f(I; \theta) \in \mathbb{R}^{2K}$ which for an image I directly regresses to a landmark vector. They minimize the mean square error over N training samples:

$$f^* = \min_f \frac{1}{N} \sum\nolimits_{n=1}^{N} \|f(I^n; \theta) - L^n\|_2. \tag{2}$$

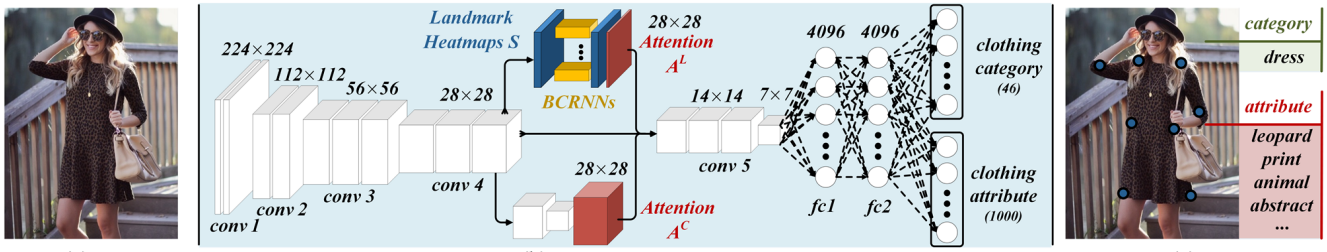

Figure 1. **Illustration of the proposed Attentive Fashion Grammar Network.** (a) Input fashion image. (b) Network architecture of our deep fashion model. A set of BCRNNs (yellow cubes) are established for capturing kinematics and symmetry grammars as global constraints for detecting clothing landmarks (blue cubes), detailed in §3.1. Fashion landmark-aware attention A^L and clothing category-driven attention A^C (red cubes) are further incorporated for enhancing clothing features and improving clothing category classification and attribute estimation (§3.2). (c) Results for clothing landmark detection, category classification and attribute estimation.

However, recent studies in pose estimation [41, 32] demonstrate this regression is highly non-linear and very difficult to learn directly, due to the fact that only one single value needs to be correctly predicted.

In this work, instead of regressing landmark positions L directly, we learn to predict a confidence map of positional distribution (*i.e.*, heatmap) for each landmark, given the input image. Let $S_k \in [0,1]^{w\times h}$ and $\hat{S}_k \in [0,1]^{w\times h}$ denote the predicted heatmap and the groundtruth heatmap (with size of $w \times h$) for the k-th landmark, respectively, our fashion network is learned as a function $f'(I;\theta') \in [0,1]^{w\times h\times K}$, via penalizing following pixel-wise mean squared differences,

$$\begin{aligned} f^* &= \min_{f'} \frac{1}{N} \sum\nolimits_{n=1}^{N} \sum\nolimits_{k=1}^{K} \mathcal{L}(f'(I^n;\theta'), L_k^n), \\ \mathcal{L}(f'(I^n;\theta'), L_k^n) &= \sum\nolimits_{u=1}^{w} \sum\nolimits_{v=1}^{h} \|S_k^n(u,v) - \hat{S}_k^n(u,v)\|_2. \end{aligned} \tag{3}$$

The groundtruth heatmap $\hat{S}_k$ is obtained by adding a 2D Gaussian filter at the groundtruth location L_k.

Fashion Grammar. We consider a total of eight landmarks (*i.e.*, K=8), namely, *left/right collar*, *left/right sleeve*, *left/right waistline*, and *left/right hem*, following previous settings [25, 26]. The natural of clothes that rich inherent structures are involved in this task, motivates us to reason the positions of landmarks in a global manner. Before going deep into our grammar network, we first detail our grammar formulations that reflect high-level knowledge of clothes. Basically, we consider the two types of fashion grammars:

• *Kinematics grammar* $\mathcal{R}^{\mathcal{K}}$ describes kinematic relations between clothing landmarks. We define 4 kinematic grammars to represent the constraints among kinematically connected clothing parts:

$$\begin{aligned} &\mathcal{R}_1^{\mathcal{K}} : \textit{l. collar} \leftrightarrow \textit{l. waistline} \leftrightarrow \textit{l. hem}, \\ &\mathcal{R}_2^{\mathcal{K}} : \textit{l. collar} \leftrightarrow \textit{l. sleeve}, \\ &\mathcal{R}_3^{\mathcal{K}} : \textit{r. collar} \leftrightarrow \textit{r. waistline} \leftrightarrow \textit{r. hem}, \\ &\mathcal{R}_4^{\mathcal{K}} : \textit{r. collar} \leftrightarrow \textit{r. sleeve}. \end{aligned} \tag{4}$$

Such grammar focuses on the clothing landmarks that connected in a human-parts kinematic chain, which satisfies human anatomical and anthropomorphic constraints.

• *Symmetry grammar* $\mathcal{R}^{\mathcal{S}}$ describes bilateral symmetric property of clothes. Symmetry of clothes is defined as the right and left sides of the cloth being mirrored reflections of each other. We consider 4 symmetric relations between clothing landmarks:

$$\begin{aligned} &\mathcal{R}_1^{\mathcal{S}} : \textit{l. collar} \leftrightarrow \textit{r. collar}, \\ &\mathcal{R}_2^{\mathcal{S}} : \textit{l. sleeve} \leftrightarrow \textit{r. sleeve}, \\ &\mathcal{R}_3^{\mathcal{S}} : \textit{l. waistline} \leftrightarrow \textit{r. waistline}, \\ &\mathcal{R}_4^{\mathcal{S}} : \textit{l. hem} \leftrightarrow \textit{r. hem}. \end{aligned} \tag{5}$$

Message Passing over Fashion Grammar. As illustrated in Fig. 2 (a), our proposed grammars upon cloth landmarks constitute a graph, where vertices specifying cloth landmark heatmaps and edges describing possible connections among vertices. To infer the optimal landmark configuration, message passing [54, 8] is favored on such loopy structures. To simulate this process, we make an approximation by performing message passing on each grammar independently and merging the output afterwards to disentangle the loopy structure.

More specifically, within the chain structure of grammar $\mathcal{R}$, the passing process is performed iteratively for each node i, consisting of two phases: the message passing phase and the readout phase. The message passing phase runs for T iterations and is defined w.r.t. message function $M(\cdot)$ and vertex update function $U(\cdot)$. In each iteration, hidden states h_i of node i is updated by computing messages coming from its neighbors j, that is,

$$\begin{aligned} m_i &\leftarrow \sum\nolimits_{j\in\mathcal{N}(i)} M(h_j), \\ h_i &\leftarrow U(m_i), \end{aligned} \tag{6}$$

where $\mathcal{N}(i)$ denotes neighbors of vertex i specified in the grammar $\mathcal{R}$.

The second phase, *i.e.*, the readout phase, infers the marginal distribution (*i.e.*, heatmaps) for each node i using h_i and readout function $\Gamma(\cdot)$, namely,

$$y_i = \Gamma(h_i). \tag{7}$$

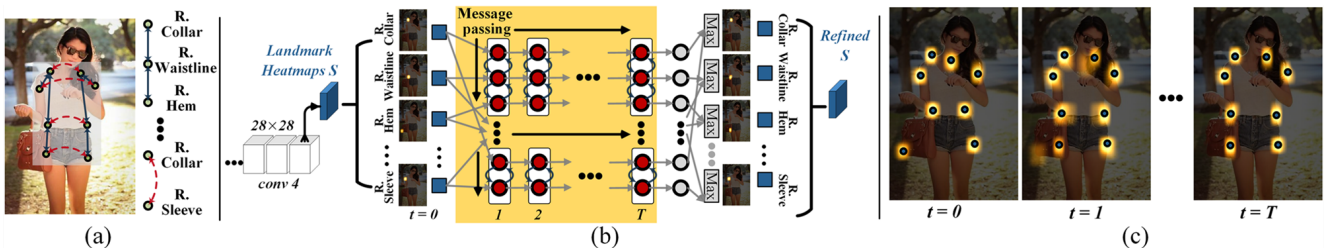

Figure 2. (a) **Illustration of our fashion grammars**, where green circles indicate groundtruth cloth landmarks, blue and red lines correspond to kinematics and symmetry grammars, respectively. (b) **Illustration of our message passing over fashion grammars**, where the blue rectangles indicate heatmaps of landmarks, and the red circles indicate BCRNN units. Within a certain BCRNN, we perform message passing over fashion grammars (one time, two directions). With stacked of BCRNNs, the messages are iteratively updated and refined landmark estimations are generated. (c) **Illustration of the refined estimations by message passing** over our fashion grammars. With the efficient message passing over grammar topology, our fashion network is able to predict more kinematically and symmetrically possible landmark layouts with high-level constraints. See §3.1 for more details.

Implementation with Recurrent Neural Network. For implementing above message passing process over grammar topology, we introduce Bidirectional Convolutional Recurrent Neural Network (BCRNN) (see Fig. 3), which is achieved by extending classical fully connected RNNs with convolution operation [36, 35].

In a high level, the bi-directionality and recurrent nature of BCRNN are favored to simulate the message passing over the grammar neighborhood system. Additionally, with the convolution operation, our model could preserve the spatial information of convolutional feature map and is able to produce pixel-wise heatmap prediction.

All the proposed grammars consist of short chain structures (*i.e.*, at most 3 vertices involved) [11], connoting that every node i in the grammar can at most have two neighbors (*i.e.*, previous node i-1 and post node i+1). Specifically, given a BCRNN, message functions $M(\cdot)$ for node i (in forward/backfoward directions) are represented as

$$\begin{aligned} m_i^f &= M^f(h_{i-1}^f) = W^f * h_{i-1}^f, \\ m_i^b &= M^b(h_{i+1}^b) = W^b * h_{i+1}^b, \end{aligned} \tag{8}$$

where $*$ denotes the convolution operator, $M^f(\cdot)$ and $M^b(\cdot)$ denote the forward and backward message function, h^f and h^b refer to the hidden states inferred from forward and backward neighbors, respectively. The hidden state h_i is thus updated accordingly

$$\begin{aligned} h_i^f &= U(m_i^f) = \tanh(m_i^f + b_h^f)\ , \\ h_i^b &= U(m_i^b) = \tanh(m_i^b + b_h^b)\ , \end{aligned} \tag{9}$$

where b_h^f and b_h^b refer to the bias term used in forward and backward inference, respectively. The readout function $\Gamma(\cdot)$ is defined as

$$y_i = \Gamma(h_i) = \sigma(W^x * x_i + h_i^f + h_i^b), \tag{10}$$

where σ is the soft-max function, x_i is the input generated by the base convolution network.

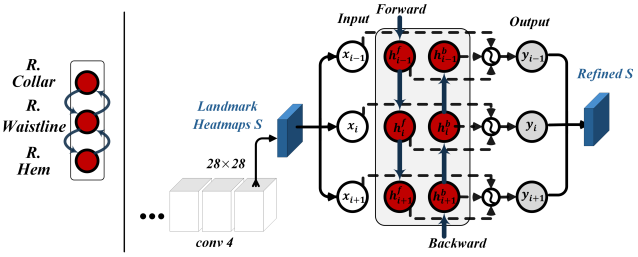

Figure 3. (a) **BCRNN with a fashion grammar**. (b) **Architecture of BCRNN**. With the input landmark predictions, the corresponding BCRNN is used for approaching massage passing over the grammar topology (a), resulting more reasonable landmark estimations. See §3.1 for more details.

We illustrate the implementation of message passing mechanism in Fig. 2 (b). By implementing massage passing with BCRNN, our network maintains the fully differentiability and obtains decent results by exchanging information along the fashion grammars.

Network Architecture. Our fashion network is based on VGG-16 architecture [39]. First, we employ features from *conv4-3* layer (the last convolution layer of the forth block) to produce K landmark heatmaps with *sigmoid* activation. Due to the max-pooling operation, we achieve $\times 8$ down-scaled heatmaps. We employ eight BCRNNs to simulate message passing procedures among cloth landmarks in each chain. A grammar BCRNN takes initial heatmaps and features from *conv4-3* as inputs and the forward process correspond to a passing process (on two directions). Generally, message passing takes several iterations to converge, while in practice three iterations are sufficient to generate satisfactory results (see more detailed discussions in §4.4). In implementation, we stack three BCRNNs (*i.e.*, T=3) for each grammar (totally 3×8 BCRNNs for all the grammars) and updated estimation via a max-pooling of predicted heatmaps from corresponding BCRNNs at the end of each stack.

3.2. Attention Modules for Clothing Category Classification

Previous studies in VQA [37, 55] and object detection [3, 45] indicate that top-down attention is good at selecting task-related locations and enhancing important features. As demonstrated in Fig. 1(b), we incorporate our fashion model with two kinds of attentions, namely fashion landmark-aware attention and category-driven attention, to improve the classification accuracy.

Fashion Landmark-Aware Attention. Clothing landmarks are keypoints centered in functional parts of clothes [25, 26]. Such representation actually provides useful information about fashion styles. Based on this observation, we introduce a landmark-aware attention mechanism that constrains our fashion model to concentrate on functional clothing regions.

For the predicted heatmaps $\{S_i\}_{i=1}^{K}$, we apply cross-channel average-pooling operation to generate a 28×28 *weight map*, A^L:

$$A^L = \frac{1}{K}\sum\nolimits_{i=1}^{K} S_i, \tag{11}$$

where $A^L \in [0,1]^{28\times 28}$. We call A^L as the landmark-aware attention. Let $F \in \mathbb{R}^{28\times 28\times 512}$ denote features obtained from *conv4-3* layer in VGG-Net, F is further updated by the landmark-aware attention A^L with same spatial dimensions, that is,

$$G_c^L = A^L \circ F_c, \quad c \in \{1,\dots,512\}, \tag{12}$$

where $\circ$ denotes the Hadamard product, F_c denotes the 2D tensor from the c-th channel of F, and G^L denotes the refined feature map. The feature is re-weighted by the landmark-aware attention and has the same size as F. Here the attention A^L works as a feature selector which produces fashion landmark aligned features. In contrast to spatial attention [20], our attention is learned in a supervised manner and encodes semantic and contextual constraints.

Clothing Category-Driven Attention. Our landmark-aware attention enhances the features from functional regions of clothes. However, such mechanism may be insufficient to discover all the informative locations to accurately classify diverse fashion categories and attributes. Inspired by recent advances in attention models [6, 42], we further propose a cloth category-driven attention A^C, which is goal directed and learned in top-down manner.

Given features F from the *conv4-3* layer, we apply a *bottom-up top-down* network [27, 30] (*e.g.*, $\times 2$ down-pooling$\rightarrow 3\times 3$ conv$\rightarrow \times 2$ down-pooling$\rightarrow 3\times 3$ conv$\rightarrow \times 4$ up-pooling) to learn a global attention map $A^C \in [0,1]^{28\times 28\times 512}$. The attention features are first pooled down to a very low resolution 7×7, then are $\times 4$ up-sampled. Thus the attention module gains a large receptive field covers all the fashion image, but is the same size as the feature map. For each position in A^C, *sigmoid* function is applied to shrink the attention values, ranging from $[0,1]$. Afterwards, we use the attention A^C to softly weight output features F:

$$G^C = A^C \circ F. \tag{13}$$

With the bottom-up top-down network, the attention obtains a large receptive field and directly enhances the task-related features from a global view. Such attention facilitates our model to learn more discriminative representations for fashion style recognition. Different from our landmark-aware attention, the category-driven attention A^C is goal-directed and learned without explicit supervision. Visualization of our attention mechanisms can be found in Fig. 4.

Network Architecture. With the feature F from the *conv4-3* layer, we consider landmark-aware attention $A^L \in [0,1]^{28\times 28}$ and clothing category-driven attention $A^C \in [0,1]^{28\times 28\times 512}$ simultaneously:

$$G_c = (1 + A^L + A_c^C) \circ F_c, \quad c \in \{1,\dots,512\}. \tag{14}$$

Such design is inspired by works in residual learning [17, 42]. If the attention models can be constructed as *identical mapping*, the performance should be no worse than its counterpart without attention. We offer more detailed analyses for our attention modules in §4.4.

As seen, the updated feature G has the same size of the feature F from *conv4-3* layer. Thus the rest layers (*pooling-4*, *conv5*s, *pooling-5*, and *fc*s) of VGG-Net can be stacked for final cloth image classification. Our attention mechanisms incorporate semantic information and global information into network and help constrain the network to focus on important clothing regions. Refined features are further used to learn classifiers on foreground clothing regions (please see Fig. 1(b)). Our whole fashion network is fully differentiable and can be trained end-to-end.

4. Experiments

In this section, we evaluate the performance of the proposed fashion model on two large-scale fashion datasets, DeepFashion: Category and Attribute Prediction Benchmark (DeepFashion-C) [25] and Fashion Landmark Dataset (FLD) [26]. Then ablation study is performed for offering more detailed exploration for the proposed approach.

4.1. Datasets

DeepFashion-C [25][1] is a large collection of $289,222$ fashion images with comprehensive annotations. Those images are collected from shopping websites and Google image search engine. Each image in this dataset is extensively labeled with 46 clothing categories, $1,000$ attributes, 8 landmarks and bounding box. The attributes are further categorized into five groups, characterizing texture, fabric,

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

Methods	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 [4]	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 [19]	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 [25]	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
Lu *et al.* [28]	86.72	92.51	-	-	-	-	-	-	-	-	-	-	-	-
Corbiere *et al.* [9]	86.30	92.80	**53.60**	63.20	39.10	48.80	50.10	59.50	38.80	48.90	30.50	38.30	23.10	30.40
Ours	**90.99**	**95.78**	50.31	**65.48**	**40.31**	48.23	**53.32**	**61.05**	40.65	**56.32**	**68.70**	**74.25**	**51.53**	**60.95**

- Detailed results are not available.

Table 1. **Quantitative results for category classification and attribute prediction on the DeepFashion-C dataset** [25]. Higher values are better. The best scores are marked in **bold**.

Methods	L.Collar	R.Collar	L.Sleeve	R.Sleeve	L.Waistline	R.Waistline	L.Hem	R.Hem	Avg.
FashionNet [25]	.0854	.0902	.0973	.0935	.0854	.0845	.0812	.0823	.0872
DFA [26]	.0628	.0637	.0658	.0621	.0726	.0702	.0658	.0663	.0660
DLAN [52]	.0570	.0611	.0672	.0647	.0703	.0694	.0624	.0627	.0643
Ours	**.0415**	**.0404**	**.0496**	**.0449**	**.0502**	**.0523**	**.0537**	**.0551**	**.0484**

Table 2. **Quantitative results for clothing landmark detection on the DeepFashion-C dataset** [25] with normalized error (NE). Lower values are better. The best scores are marked in **bold**.

shape, part, and style, respectively. Based on this dataset, we extensively examine the performance of our deep fashion model in fashion landmark detection, clothing category and attribute classification.

FLD [26][2] is collected for fashion landmark detection. It contains $123,016$ clothing images, with diverse and large pose/zoom-in variations. For each image, the annotations for 8 fashion landmarks are offered. In our experiments, we use this dataset to only evaluate fashion landmark detection, as no garment category annotations are provided.

4.2. Experiments on DeepFashion-C Dataset

Experimental Setup. We follow the settings in DeepFashion-C [25] for training and testing. More specifically, $209,222$ fashion images are used for training and $40,000$ images are used for validation. The evaluation is performed on the remaining $40,000$ images. For training and testing, following [25, 28], we crop each image using ground truth bounding box. 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 [25], which is obtained by ranking the $1,000$ classification scores and determine how many attributes have been matched in the top-*k* list. For clothing fashion detection, we adopt normalized error (NE) metric [26] for evaluation. NE refers to the ℓ_2 distance between predicted landmarks and groundtruth in the normalized coordinate space (*i.e.*, normalized with respect to the width/height of the image).

Implementation Details. Our network is built upon VGG-16 with fashion grammar BCRNNs (§3.1) and attention modules (§3.2). We resize all the cropped images into 224×224. Thus our network would generate eight 28×28 heatmaps for clothing landmarks. We replace the last fully connected layer by two branched fully connected layers for fashion category classification and attribute estimation. In DeepFashion-C dataset, each image receives one category label and multiple attribute labels (average 3.3 attributes per image). For category classification, we apply 1-of-K softmax loss for training the branch of fashion category. For training the other branch of attribute prediction, we apply asymmetric weighted cross-entropy loss [29], due to the data unbalance between positive and negative samples.

Our model is implemented in Python with the help of TensorFlow backend, and trained with Adam optimizer. For the BCRNNs and category-related attention module, we use 3×3 kernel for all the convolution operations. In each training iteration, we use a mini-batch of 10 images, which are randomly sampled from DeepFashion-C dataset. We first pre-train the former four convolution blocks of our network with cloth landmark detection with two epochs. Then our whole model is trained with ten epochs. The learning rate is set as 0.0001 and is decreased by a factor of 10 every two epochs. We perform early-stopping without improvements on the validation set. The entire training procedure takes about 40 hours with a single NVIDIA TITAN X GPU and a 4.0 GHz Intel processor with 32GB memory.

Performance Evaluation. For category classification and attribute prediction, we compare our method with five recent deep learning models [4, 19, 25, 28, 9] that showed compelling performance in clothes recognition and human attribute classification. For cloth landmark detection, we compare our model with three top-performing deep learning models [25, 26, 52]. Note that the results are biased

[2]This dataset is available at: http://mmlab.ie.cuhk.edu.hk/projects/DeepFashion/LandmarkDetection.html

Methods	L.Collar	R.Collar	L.Sleeve	R.Sleeve	L.Waistline	R.Waistline	L.Hem	R.Hem	Avg.
FashionNet [25]	.0784	.0803	.0975	.0923	.0874	.0821	.0802	.0893	.0859
DFA [26]	.048	.048	.091	.089	-	-	.071	.072	.068
DLAN [52]	.0531	.0547	.0705	.0735	.0752	.0748	.0693	.0675	.0672
Ours	**.0463**	**.0471**	**.0627**	**.0614**	**.0635**	**.0692**	**.0635**	**.0527**	**.0583**

- Detailed results are not released.

Table 3. **Quantitative results for clothing landmark detection on the FLD dataset** [26] with normalized error (NE). Lower values are better. The best scores are marked in **bold**.

Figure 4. **Clothing category classification results and visualization of attention mechanisms** on DeepFashion-C dataset [25]. The correct predictions are marked in green and the wrong predications are marked in red. Best viewed in color. For category-aware attention, we randomly select attentions from 2 channels for visualization.

towards [25], as it is pre-trained with $300,000$ images from DeepFashion and fine-tuned on the DeepFashion-C. For the model [26], the training settings follow the standard protocol in DeepFashion-C. For unconstrained landmark detection model [52], which is reimplemented according to the authors's descriptions, we use the cropped fashion images as inputs for the sake of fair comparison.

Table 1 summarizes the performance of different methods on clothing category classification and attribute prediction. As seen, the proposed fashion model achieves the best score on clothing category classification (top-3: 90.99, top-5: 95.78) and the best average score over all attributes (top-3: 51.53, top-5: 60.95). In Table 2 we present comparison results with other models [25, 26, 52] for clothing landmark detection. Our total NE score achieves state-of-the-art at 0.0484, which is much lower than the closest competitor (0.0643), and it is noteworthy that our method consistently improves the accuracy in all landmarks.

4.3. Experiments on FLD Dataset

Experimental Setup. FLD dataset [26] is specially designed for fashion landmark detection. Each image in this dataset is labeled with eight landmarks. With dataset, we study the performance of deep fashion model on fashion landmark detection. Following the protocol in FLD, $83,033$ images and $19,992$ fashion images are used for training and validating, $19,991$ images are used for testing. NE metric suggested by FLD is used for evaluation. The images are also cropped according to the available bounding boxes.

Implementation Details. Since we only concentrate on fashion landmark detection. We preserve the former four convolution blocks (without *pooling4*) and our fashion grammar BCRNNs, which are used for estimating heatmaps for landmarks. 3×3 convolution kernels are also used in BCRNNs. Other settings are similar to the ones used for DeepFashion-C dataset in § 4.2.

Performance Evaluation. We compare our model with FashionNet [25], DFA [26] and DLAN [52]. For sake of fair comparison, we train FashionNet [25] and DLAN [52] following standard train/val/test settings in FLD. For DFA, we preserve their original results reported in [26]. But the results are biased for DFA, since it's trained with extra clothing labels (upper-/lower-/whole-body clothes).

In Table 3, we report the comparison results on the FLD dataset with NE score. Our model again achieves state-of-the-art at 0.0583 and consistently outperforms other competitors on all of the fashion landmarks. Note that our method achieves such high accuracy without any preprocessing (*e.g.*, [26] groups cloth images into different clusters and considers extra clothing labels). Sampled landmark detection results are presented in Fig. 5.

4.4. Ablation Study

In this section, we perform an in-depth study of each component in our deep fashion network.

Effectiveness of Fashion Grammars and Message

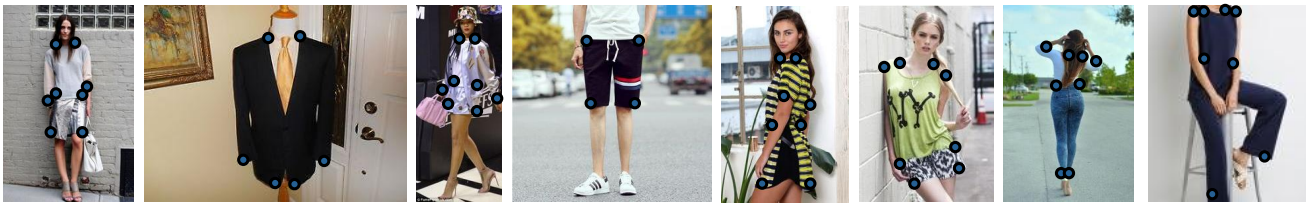

Figure 5. **Visual results for clothing landmark detection** on DeepFashion-C [25] (first row) and FLD [26] (bottom row). The detected landmarks are marked in blue circles. Best viewed in color.

Variants	DeepFashion-C		FLD	
	NE ↓	ΔNE ↓	NE ↓	ΔNE ↓
Ours (iteration 3)	**.0484**	-	**.0583**	-
Ours w/o $\mathcal{R}^K$	.0525	.0041	.0659	.0076
Ours w/o $\mathcal{R}^S$	.0538	.0054	.0641	.0058
Ours w/o $\mathcal{R}^K$ & $\mathcal{R}^S$	.0615	.0131	.0681	.0098
Ours-iteration 1	.0579	.0095	.0657	.0074
Ours-iteration 2	.0512	.0028	.0632	.0049

Table 4. **Ablation study for the effect of fashion grammars and message passing** on DeepFashion-C [25] and FLD [26] datasets.

Passing. We first examine the effectiveness of our fashion grammars, which are models via BCRNNs. In §3.1, we consider two types of grammars that account for kinematic dependencies $\mathcal{R}^K$ and symmetric relations $\mathcal{R}^S$, respectively. Three baselines are considered:

• *Ours w/o* $\mathcal{R}^K$*:* training our model without considering kinematics grammar $\mathcal{R}^K$.

• *Ours w/o* $\mathcal{R}^S$*:* training our model without considering symmetry grammar $\mathcal{R}^K$.

• *Ours w/o* $\mathcal{R}^K$ *&*$\mathcal{R}^S$*:* training our model without considering kinematics grammar $\mathcal{R}^K$ and symmetry grammar $\mathcal{R}^K$.

For accessing the effect of iterative message passing over grammars, we report two baselines: *Ours-iteration 1*, *Ours-iteration 2*, which correspond to the results from different passing iterations. The final results (baseline *Ours*) can be viewed as the results in the third passing iteration.

We carry out experiments on the DeepFashion-C [25] and FLD [26] datasets with landmark detection task, and measure the performance using normalized error (NE). Table 4 shows the performance of each of the baselines described above. We can observe that fashion grammars provides domain-specific knowledge for regularizing the landmark outputs, boosting further the results ($0.0615 \rightarrow 0.0484$ on DeepFashion-C, $0.0681 \rightarrow 0.0583$ on FLD). In addition, both kinematics and symmetry grammars contribute the improvement. We also observe the massage passing is able to gradually improve the performance.

Effectiveness of Attention Mechanisms. Next we study the influence of our attention modules. In §3.2, we consider two kinds of attentions, namely landmark-aware attention A^L and cloth category-driven attention A^C, for enhancing landmark-aligned and category-related features. Three variants derived from our method are considered:

Variants	Category		Attribute	
	top-3 ↑	top-5 ↑	top-3 ↑	top-5 ↑
Ours (w/ A^L & A^C)	**90.99**	**95.78**	**51.53**	**60.95**
Ours w/o A^L	85.27	91.32	48.29	56.65
Ours w/o A^C	87.75	93.67	49.93	58.78
Ours w/o A^L & A^C	83.23	89.51	43.28	53.54

Table 5. **Ablation study for the effectiveness of attention mechanisms** on DeepFashion-C [25] dataset.

• *Ours w/o* A^L*:* training our model without considering landmark-aware attention A^L.

• *Ours w/o* A^C*:* training our model without considering cloth category-driven attention A^C.

• *Ours w/o* A^L *&* A^C*:* training our model without considering landmark-aware attention A^L and cloth category-driven attention A^C.

We experiment on the DeepFashion-C dataset with tasks of cloth category classification and fashion attribute estimation, and measure the performance using the top-*k* accuracy and top-*k* recall. As evident in Table 5, by disabling attentions A^C and A^L, we observe significant drop of performance, on both tasks. This suggests that our attention models indeed improve the discriminability of deep learning features. When enabling A^C or A^L attention module, we can achieve better performance. The best performance is achieved via combining A^C and A^L.

5. Conclusions

In this paper, we proposed a knowledge-driven and attention-involved fashion model. Our model extended neural network with domain-specific grammars, learning to a powerful fashion network that inherits the advantages of both. In our fashion grammar representations, kinetic dependencies and symmetric relations are encoded. We introduce Bidirectional Convolutional Recurrent Neural Networks (BCRNNs) for modeling the message passing over our grammar topologies, leading to a fully differentiable network that can be end-to-end training. We further introduced two types of attentions for improving the performance of clothing image classification. We demonstrate our model on two benchmarks, and achieve the state-of-the-art fashion image classification and landmark detection performance against recent methods.

References

[1] Z. Al-Halah, R. Stiefelhagen, and K. Grauman. Fashion forward: Forecasting visual style in fashion. In *ICCV*, 2017. 2

[2] Y. Amit and A. Trouvé. Pop: Patchwork of parts models for object recognition. *IJCV*, 75(2):267, 2007. 2

[3] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu, et al. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *ICCV*, 2015. 2, 5

[4] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. *ECCV*, 2012. 2, 6

[5] H. Chen, Z. J. Xu, Z. Q. Liu, and S.-C. Zhu. Composite templates for cloth modeling and sketching. In *CVPR*, 2006. 2

[6] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, 2016. 2, 5

[7] 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*, 2015. 2

[8] X. Chu, W. Ouyang, H. Li, and X. Wang. CRF-CNN: Modeling structured information in human pose estimation. In *NIPS*, 2016. 3

[9] C. Corbiere, H. Ben-Younes, A. Rame, and C. Ollion. Leveraging weakly annotated data for fashion image retrieval and label prediction. In *ICCV workshop*, 2017. 1, 2, 6

[10] H.-S. Fang, Y. Xu, W. Wang, X. Liu, and S.-C. Zhu. Learning pose grammar to encode human body configuration for 3D pose estimation. In *AAAI*, 2018. 2

[11] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. Neural message passing for quantum chemistry. In *ICML*, 2017. 4

[12] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, 2017. 2

[13] M. Hadi 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

[14] F. Han and S.-C. Zhu. Bottom-up/top-down image parsing with attribute grammar. *IEEE TPAMI*, 31(1):59–73, 2009. 2

[15] X. Han, Z. Wu, P. X. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, and L. S. Davis. Automatic spatially-aware fashion concept discovery. In *ICCV*, 2017. 2

[16] X. Han, Z. Wu, Y.-G. Jiang, and L. S. Davis. Learning fashion compatibility with bidirectional lstms. In *ACM MM*, 2017. 2

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[18] W.-L. Hsiao and K. Grauman. Learning the latent "look": Unsupervised discovery of a style-coherent embedding from fashion images. In *ICCV*, 2017. 2

[19] 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, 6

[20] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. In *NIPS*, 2015. 2, 5

[21] M. H. Kiapour, K. Yamaguchi, A. C. Berg, and T. L. Berg. Hipster wars: Discovering elements of fashion styles. In *ECCV*, 2014. 2

[22] X. Liang, C. Xu, X. Shen, J. Yang, S. Liu, J. Tang, L. Lin, and S. Yan. Human parsing with contextualized convolutional neural network. In *ICCV*, 2015. 1, 2

[23] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *CVPR*, 2012. 2

[24] T. Liu, S. Chaudhuri, V. G. Kim, Q. Huang, N. J. Mitra, and T. Funkhouser. Creating consistent scene graphs using a probabilistic grammar. *TOG*, 33(6):211, 2014. 2

[25] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 1, 2, 3, 5, 6, 7, 8

[26] Z. Liu, S. Yan, P. Luo, X. Wang, and X. Tang. Fashion landmark detection in the wild. In *ECCV*, 2016. 2, 3, 5, 6, 7, 8

[27] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 5

[28] Y. Lu, A. Kumar, S. Zhai, Y. Cheng, T. Javidi, and R. Feris. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. *CVPR*, 2017. 1, 2, 6

[29] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich. Feedforward semantic segmentation with zoom-out features. In *CVPR*, 2015. 6

[30] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 5

[31] S. Park, X. Nie, and S.-C. Zhu. Attribute and-or grammar for joint parsing of human pose, parts and attributes. *IEEE TPAMI*, 2017. 2

[32] T. Pfister, J. Charles, and A. Zisserman. Flowing convnets for human pose estimation in videos. In *ICCV*, 2015. 3

[33] S. Qi, S. Huang, P. Wei, and S.-C. Zhu. Predicting human activities using stochastic grammar. In *ICCV*, 2017. 2

[34] S. Qi, Y. Zhu, S. Huang, C. Jiang, and S.-C. Zhu. Human-centric indoor scene synthesis using stochastic grammar. In *CVPR*, 2018. 2

[35] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997. 4

[36] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *NIPS*, 2015. 4

[37] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. In *CVPR*, 2016. 2, 5

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

[39] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 4

[40] Z. Song, M. Wang, X.-S. Hua, and S. Yan. Predicting occupation via human clothing and contexts. In *ICCV*, 2011. 2

[41] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, 2014. 3

[42] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. In *CVPR*, 2017. 2, 5

[43] W. Wang and J. Shen. Deep visual attention prediction. *IEEE TIP*, 27(5):2368–2378, 2018. 2

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

[45] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *CVPR*, 2015. 2, 5

[46] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 2

[47] Y. Xu, L. Lin, W.-S. Zheng, and X. Liu. Human re-identification by matching compositional template with cluster sampling. In *ICCV*, 2013. 2

[48] Y. Xu, X. Liu, Y. Liu, and S.-C. Zhu. Multi-view people tracking via hierarchical trajectory composition. In *CVPR*, 2016. 2

[49] Y. Xu, X. Liu, L. Qin, and S.-C. Zhu. Cross-view people tracking by scene-centered spatio-temporal parsing. In *AAAI*, 2017. 2

[50] K. Yamaguchi, M. Hadi Kiapour, and T. L. Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *ICCV*, 2013. 2

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

[52] S. Yan, Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Unconstrained fashion landmark detection via hierarchical recurrent transformer networks. In *ACM MM*, 2017. 2, 6, 7

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

[54] W. Yang, W. Ouyang, H. Li, and X. Wang. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In *CVPR*, 2016. 3

[55] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *CVPR*, 2016. 2, 5

[56] Y. Zhao and S.-C. Zhu. Image parsing with stochastic scene grammar. In *NIPS*, 2011. 2