Clothes Co-Parsing Via Joint Image Segmentation and Labeling With Application to Clothing Retrieval

Xiaodan Liang, Liang Lin, Wei Yang, Ping Luo, Junshi Huang, and Shuicheng Yan

***Abstract*—This paper aims at developing an integrated system for clothing co-parsing (CCP), in order to jointly parse a set of clothing images (unsegmented but annotated with tags) into semantic configurations. A novel data-driven system consisting of two phases of inference is proposed. The first phase, referred as "image cosegmentation," iterates to extract consistent regions on images and jointly refines the regions over all images by employing the exemplar-SVM technique [1]. In the second phase (i.e., "region colabeling"), we construct a multiimage graphical model by taking the segmented regions as vertices, and incorporating several contexts of clothing configuration (e.g., item locations and mutual interactions). The joint label assignment can be solved using the efficient Graph Cuts algorithm. In addition to evaluate our framework on the Fashionista dataset [2], we construct a dataset called the SYSU-Clothes dataset consisting of 2098 high-resolution street fashion photos to demonstrate the performance of our system. We achieve 90.29%/88.23% segmentation accuracy and 65.52%/63.89% recognition rate on the Fashionista and the SYSU-Clothes datasets, respectively, which are superior compared with the previous methods. Furthermore, we apply our method on a challenging task, i.e., cross-domain clothing retrieval: given user photo depicting a clothing image, retrieving the same clothing items from online shopping stores based on the fine-grained parsing results.**

***Index Terms*—Clothes recognition, fashion understanding, human-centric computing, image parsing.**

I. Introduction

CLOTHING recognition and parsing have huge potentials in Internet-based e-commerce, as the revenue of online clothing sale keeps highly increasing every year. The related techniques also benefit a wide range of human centric real applications. For example, clothing parsing, i.e., segmenting and labeling clothing items, can be utilized as one of the key components of virtual clothing try-on system [3] that simulates the selected virtual clothing on user's body in real-time. The applications of clothing retrieval [4], and outfit recommendation [5]–[7] also strongly desire the fine-grained clothing parsing and interpretation. Moreover, clothing recognition can be also employed for person re-identification [8], especially under the surveillance scenarios where identifiable faces are unavailable, since the good interpretation of one's clothing may provide extra useful cues to verify a person.

Despite several interesting works [2], [9]–[12] have been proposed on this task and shown promising results, clothing parsing has not been fully solved, especially for real-world photos with large amounts of diverse clothing tags and dressing styles. Specifically, the difficulties lie in the following aspects. First, existing clothing parsing systems usually employ supervised learning to assign semantic tags to all pixels within an image, so that abundant pixel annotation of clothing items are required. This annotation, however, often costs expensively and processes inefficiently. Second, the appearances of clothing and garment items have much larger variations due to different styles, textures and materials, compared with other common objects. Third, severe occlusions between clothing items and human bodies often exist in unconstrained environments, as well as complex human poses and self-occlusions. Finally, the number of fine-grained clothing tags is very large, e.g., Fashionista dataset [2] contains more than 50 tags of clothing items. In contrast, existing co-segmentation systems [13], [14] have been developed to deal with much fewer the semantic tags. These challenges thus limit the clothing parsing performance of applying traditional object recognition or semantic segmentation approaches.

The explosive development of social networks, photo sharing and e-commerce websites provides possible access to large amounts of fashion photos and user data, and further help associate the clothing images with clothing tags [2]–[9]. Under this background, an interesting problem arises: Is it possible to automatically transfer fine-grained clothing tags at image level to the regions or pixels? To answer this question, in this paper, we develop an engineered clothes co-parsing (CCP) framework[1] to jointly parse a batch of clothing images and produce accurate pixelwise annotation of clothing items. Our system consists of two sequential phases of inference over a set of clothing images, i.e., image co-segmentation for extracting distinguishable clothing regions, and region co-labeling for recognizing various garment items, as illustrated in Fig. 1. Furthermore, the con-

Manuscript received July 24, 2015; revised November 25, 2015 and February 03, 2016; accepted March 14, 2016. Date of publication March 16, 2016; date of current version May 13, 2016. This work was supported in part by the Guangdong Natural Science Foundation under Grant S2013050014548 and Grant 2014A030313201, in part by Program of Guangzhou Zhujiang Star of Science and Technology under Grant 2013J2200067, and in part by Guangdong Science and Technology Program under Grant 2015B010128009. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sen-Ching Samson Cheung. *(Corresponding author: Liang Lin.)*

X. Liang and L. Lin are with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China (e-mail: xdliang328@gmail.com; linliang@ieee.org).

W. Yang and P. Luo are with the Chinese University of Hong Kong, Hong Kong, China (e-mail: platero.yang@gmail.com; pluo.lhi@gmail.com).

J. Huang and S. Yan are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119077 (e-mail: junshi.huang@gmail.com; eleyans@nus.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2016.2542983

[1][Online]. Available: http://vision.sysu.edu.cn/projects/clothing-co-parsing/

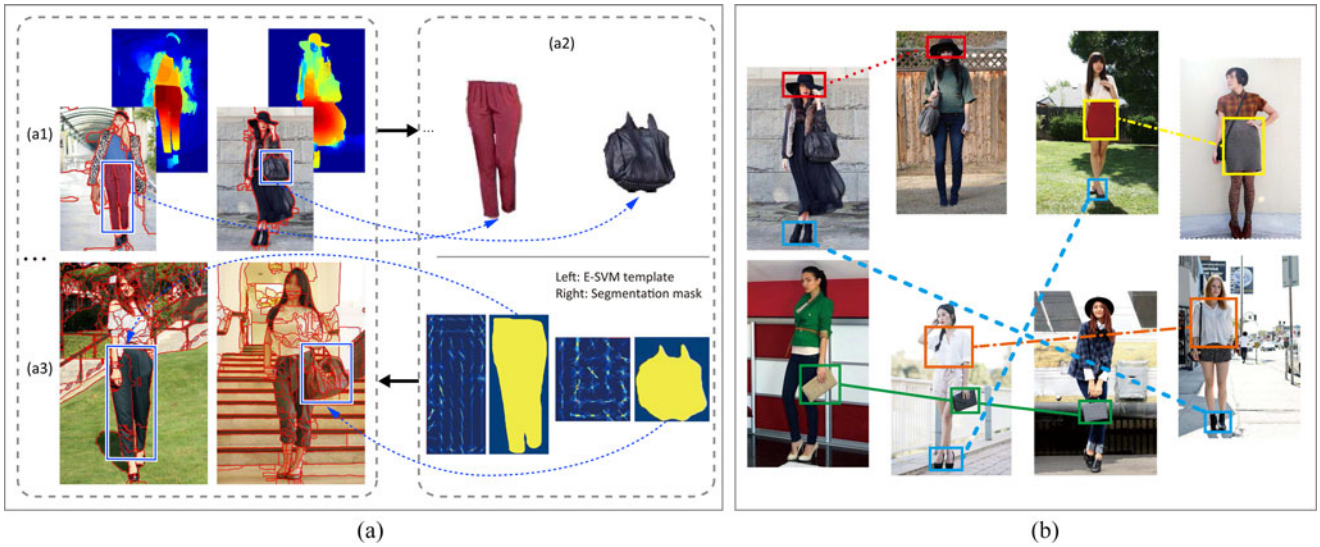


Fig. 1. Illustration of the proposed clothes co-parsing framework, which consists of two sequential phases of optimization: (a) clothing co-segmentation for extracting coherent clothes regions, and (b) region co-labeling for recognizing various clothes garments. Specifically, clothing co-segmentation iterates with three steps: (a1) grouping superpixels into regions and detecting salient regions, (a2) selecting confident foreground regions according to the saliency maps to train E-SVM classifiers, and (a3) propagating segmentations by applying E-SVM templates over all images. Given the segmented regions, clothing co-labeling is achieved based on a multi-image graphical model, as illustrated in (b).

texts of clothing configuration are also exploited, e.g., spatial locations and mutual relations of clothing items, inspired by the successes of object/scene context modeling [15]–[17]. In the following, we briefly discuss the motivations and main components of our framework.

At the first phase (i.e., image co-segmentation), our method iteratively refines the regions grouped over all images by utilizing the exemplar-SVM (E-SVM) technique [1]. At the beginning, the superpixels for each image are extracted and then grouped into regions, where most regions are often cluttered and meaningless due to the diversity of clothing appearances and human body variations, as shown in Fig. 1(a1). However, some coherent regions [in Fig. 1(a2)] which satisfy some certain criteria (e.g., size and location constraints), can be still selected. Then, several E-SVM classifiers are trained for the selected regions using the HOG feature, i.e., one classifier for one region, and then a set of region-based detectors are generated, as shown in Fig. 1(a3), which are employed as the top-down templates to localize similar regions over all images. In this way, segmentations are refined jointly, as more coherent regions are generated by the trained E-SVM classifiers. This process is inspired by the observation that clothing items of the same fine-grained category often share similar patterns (i.e., shapes and structures). In the literature, Kuettel *et al.* [18] also proposed to propagate segmentations through HOG-based matching.

Unlike the traditional approaches that perform supervised learning to predict the labels of segmented regions of all images, we design the second phase (i.e., region co-labeling) in a data-driven manner. A multi-image graphical model is constructed by taking the regions as vertices of graph, inspired by [19]. In our graphical model, the adjacent regions within each image as well as regions across different images are linked, which share similar appearance and latent semantic tags. Thus we can borrow statistical strength from similar regions in different images and assign labels jointly, as Fig. 1(b) illustrates. The efficient Graph Cuts algorithm [20] is finally utilized for co-labeling optimization that incorporates several contextual constraints defined on the clothing items.

Moreover, a large-scale clothing parsing dataset annotated with pixel-wise labeling is proposed for evaluating clothing co-parsing, which includes more realistic and general challenges, e.g., disordered backgrounds and multiform human poses, compared with the existing clothing datasets [2], [21], [22], and more fine-grained clothing tags, compared with the recent clothing dataset [23], [24]. We demonstrate promising performances and applicable potentials of our system in the experiments.

The cross-domain clothing retrieval plays an important role in human-centric applications, such as mobile product fashion analysis [25], [26] and person re-identification [8]. To further validate the effectiveness of our method, the cross-domain clothing retrieval based on the clothing parsing results is also explored in this paper. The fine-grained semantic regions parsed by our method enable precisely locating the region of interest for the query and gallery images while the previous clothing retrieval methods [4] just used the whole image or clothing detection results as the query. Incorporating clothe parsing into clothing retrieval can reduce the effect of the possible background clutters included in the photo that may lead to unsatisfied retrieval results. We collect a large dataset composed of cross-scenario image pairs, which includes about 10 000 online (product photos from merchants) and offline (photos uploaded by customers) image pairs, as shown in Fig. 3. Each image has 124 fine-grained semantic attribute types. Given an offline clothing image from the "street" domain as the query, we first parse this query image into semantic regions of clothing items and then retrieve the

same or similar clothing items from a large-scale gallery of professional online shopping images. Based on the annotated fine-grained clothing attribute types, the domain-specific neural network is finetuned based on the pre-trained convolutional neural network (CNN) [27] for generating more effective feature representation. The top-20 retrieval accuracy is improved by 4.9% when using the proposed method as the basic clothing parser other than using the method [2].

The rest of this paper is organized as follows. Section I-A presents a brief review of related work. We introduce the probabilistic formulation of our framework in Section II, and then discuss the implementation of the two phases in Section III. The cross-domain clothing retrieval based on clothing parsing is described in Section IV. The experiments and comparisons are presented in Section V, and the paper concludes in Section VI.

A. Related Work

Currently, existing efforts on clothing/human segmentation and recognition mainly focused on constructing expressive models to address various clothing styles and appearances [2], [22], [28]–[33]. One classic work [22] proposed a composite And-Or graph template for modeling and parsing clothing configurations. Later works studied on blocking models for highly occluded group images [30], or deformable spatial priors modeling to improve performance of clothing parsing [28]. Recent approaches incorporated shape-based human model [29], or pose estimation and supervised region labeling [2], and achieved impressive results. For example, a shape-based human model was presented in [29] for human/clothing segmentation, which assembled candidate parts from an over-segmentation of the image and matched them to a library of shape exemplars. Yamaguchi *et al.* [2] demonstrated very impressive results on pixel-wise clothing parsing by integrating pose estimation and supervised region labeling. In addition, Liu *et al.* [34] proposed to utilize the user-generated color-category as the weak supervision for clothe parsing. And we believe the effort of our system will further enhance their performances by providing reliable annotations in an easy way.

Clothes co-parsing is also highly related to image/object co-labeling, where a batch of input images containing similar objects are processed jointly [35]–[37]. For example, unsupervised shape guided approaches were adopted in [38] to achieve single object category co-labeling. Winn *et al.* [39] incorporated automatic image segmentation and spatially coherent latent topic model to obtain unsupervised multi-class image labeling. These methods, however, solved the problem in an unsupervised manner, and might be intractable under circumstances with large numbers of categories and diverse appearances. To deal with more complex scenario, some recent works focused on supervised label propagation, utilizing pixelwise label map in the training set and propagating labels to unseen images. Pioneering work of Liu *et al.* [35] proposed to propagate labels over scene images using a bi-layer sparse coding (BSC) formulation.

Deep CNNs have achieved significant successes in many areas of computer vision [8],[23]. Recently, Liang *et al.* [23] proposed to use two separate convolutional networks to predict the template coefficients for label masks and the corresponding locations, respectively. The quasi-parametric human parsing model [24] has also been proposed to predict the matching confidence and displacement of the best matched region for a particular semantic region in one KNN image. These works have conducted the experiments on the dataset with much fewer tags (i.e., 18 tags) and required large-scale pixel-wise annotations for network training. Besides, some interesting yet commercially valuable applications such as clothing retrieval and recommendation systems have also been proposed. Prior approaches for clothing retrieval based on deep learning have outperformed previous methods based on traditional image representations [40]. Several domain adaptation methods based on deep learning have been recently proposed [4]. For example, Huang *et al.* [4] proposed a Dual Attribute-aware Ranking Network for retrieval feature learning. Different from all these works, our retrieval system is capitalized on the clothing parsing results by the proposed framework instead of a simple whole image. This paper is an extension of our previous conference paper [41].

II. Probabilistic Framework

In this paper, the CCP task is formulated as a probabilistic model. Let $\mathbf{I} = \{I_i\}_{i=1}^N$ denote a set of clothing images with tags $\{T_i\}_{i=1}^N$. Each image I is represented by a set of superpixels $I = \{s_j\}_{j=1}^M$, which can be further grouped into a set of coherent regions by utilizing the segmentation propagation. Four additional variables are associated with each image I:

1) the regions $\{r_k\}_{k=1}^K$, each of which consists of a set of superpixels;
2) the garment label for each region: $\ell_k \in T, k = 1, \ldots, K$;
3) the E-SVM weights w_k trained for each selected region; and
4) the segmentation propagations $C = (x, y, m)$, where (x, y) is the location and m is the segmentation mask of an E-SVM, indicating segmentation mask m can be propagated to the position (x, y) of I, as illustrated in Fig. 1(a).

Let $\mathbf{R} = \{R_i = \{r_{ik}\}\}$, $\mathbf{L} = \{L_i = \{\ell_{ik}\}\}$, $\mathbf{W} = \{W_i = \{w_{ik}\}\}$ and $\mathbf{C} = \{C_i\}$. The model parameters can be optimized by maximizing the following posterior probability:

$$\{\mathbf{L}^*, \mathbf{R}^*, \mathbf{W}^*, \mathbf{C}^*\} = \arg\max P(\mathbf{L}, \mathbf{R}, \mathbf{W}, \mathbf{C}|\mathbf{I}) \quad (1)$$

which can be factorized by

$$P(\mathbf{L}, \mathbf{R}, \mathbf{W}, \mathbf{C}\}|\mathbf{I}) \propto \overbrace{P(\mathbf{L}|\mathbf{R}, \mathbf{C})}^{co-labeling} \times \overbrace{\prod_{i=1}^{N} P(R_i|C_i, I_i)P(W_i|R_i) \times P(C_i|\mathbf{W}, I_i)}^{co-segmentation}. \quad (2)$$

The optimization of (2) is performed with two phases: 1) clothing image co-segmentation and 2) region co-labeling.

In *phase* (I), the optimal regions are obtained by maximizing $P(R|C,I)$ in (2). We denote the superpixel grouping indicator as $o_j \in \{1,\ldots,K\}$, which represents which of the K regions the superpixel s_j belongs. Then each region can be represented as several superpixels, as $r_k = \{s_j | o_j = k\}$. Given the current segmentation propagation C, $P(R|C,I)$ can be calculated as

$$P(R|C,I) = \prod_k P(r_k|C,I) \propto \prod_j P(o_j|C,I) \propto \prod_{j=1}^{M} P(o_j,s_j) \prod_{mn} P(o_m,o_n,s_m,s_n|C) \quad (3)$$

where the unary potential $P(o_j,s_j) \propto \exp\{-d(s_j,o_j)\}$ indicates the probability of superpixel s_j belonging to a region, $d(s_j,o_j)$ represents the spatial displacement between s_j and its corresponding region. $P(o_m,o_n,s_m,s_n|C)$ is the pairwise potential function, which encourages smoothness between neighboring superpixels.

After grouping superpixels into regions, several coherent regions can be selected to train an ensemble of E-SVMs, by maximizing $P(W|R)$ defined as follows:

$$P(W|R) = \prod_k P(w_k|r_k) \propto \prod_k \exp\{-E(w_k,r_k)\cdot\phi(r_j)\} \quad (4)$$

where $\phi(r_j)$ is an indicator exihibiting whether r_j has been chosen for training E-SVM. $E(w_k,r_k)$ is the convex energy function of E-SVM.

Finally, $P(C_i|\mathbf{W},I_i)$ in (2) is defined based on the responses of E-SVM classifiers. The $P(C_i|\mathbf{W},I_i)$ is maximized by selecting the top k detections of each E-SVM classifier as the segmentation propagations by the sliding window scheme.

In *phase* (II), we assign a garment tag to each region by modeling the region co-labeling problem as the optimization of a multi-image graphical model

$$P(\mathbf{L}|\mathbf{R},\mathbf{C}) \propto \prod_i^N \prod_k^K P(\ell_{ik},r_{ik}) \cdot \prod_{mn} P(\ell_m,\ell_n,r_m,r_n) \prod_{uv} Q(\ell_u,\ell_v,r_u,r_v|\mathbf{C}) \quad (5)$$

where $P(\ell_{ik},r_{ik})$ indicates the singleton potential of assigning label ℓ_{ik} to region r_{ik}, and $P(\ell_m,\ell_n,r_m,r_n)$ the interior affinity model capturing compatibility among regions within one image, and $Q(\ell_u,\ell_v,r_u,r_v|\mathbf{C})$ the exterior affinity model for regions belonging to different images, in which r_u and r_v are connected under the segmentation propagation $\mathbf{C}$. More details are discussed in Section III-B.

III. Co-Segmentation

In this section, the two phases of CCP as well as their implementation details are described. The overall procedure is outlined in Algorithm 1.

Algorithm 1: The Sketch of Clothes Co-parsing.

Input:
A set of clothing images $\mathbf{I} = \{I_i\}_{i=1}^N$ with tags $\{T_i\}_{i=1}^N$.

Output:
The segmented regions $\mathbf{R}$ with their corresponding labels $\mathbf{L}$.

PHASE (I): Image Co-Segmentation
Initialize the segmentation propagation C as the whole image.

Repeat
1 For each image I, group its superpixels into regions R under the guidance of the segmentation propagations C by maximizing $P(R|C,I)$ in (3);
2 Train E-SVM parameters for each selected region by minimizing the energy in (7).
3 Propagate segmentations across images by detections from the trained E-SVM classifiers by (8).
Until Regions are not changed during the last iteration

PHASE (II): Contextualized Co-Labeling
1 Construct the multi-image graphical model;
2 Solving the optimal label assignment $\mathbf{L}^*$ by optimizing the probability defined on the graphical model as in (5) by Graph Cuts.

A. Unsupervised Image Co-Segmentation

The optimization in the co-segmentation is to estimate a variable while keeping others fixed, e.g., estimating R, with W,C fixed. Thus the first phase iterates across the whole dataset (including both training and testing images) between three steps, as follows.

1) Superpixel Grouping: The MRF model defined in (3) is the standard pipeline for superpixel grouping. However, the number of regions need to be specified, which is not an applicable assumption of our problem, since the number of garment tags does not strictly correspond to the number of semantic regions.
To automatically determine the number of semantic regions, the superpixel indicator o_j is replaced by a list of *binary* variables o^e defined on the edges between the neighboring superpixels. Let e denote an edge, $o^e = 1$ if two superpixels s_1^e and s_2^e connected by e belong to the same semantic region, otherwise $o^e = 0$. The binary variable o^c with $o^c = 1$ is denoted to indicate all the superpixels within the mask of the segmentation propagation c belonging to the same semantic region, otherwise $o^c = 0$. Then maximizing (3) can be equivalently optimized by the following linear programming problem:

$$\arg\min_{o^e,o^c} \sum_e d(s_1^e,s_2^e)\cdot o^e - \sum_{c\in C} h(\{s_j|s_j\subset c\})\cdot o^c \quad (6)$$

where $d(s_1^e,s_2^e)$ indicates the dissimilarity between two superpixels, and $h(\cdot)$ measures the consistence of grouping all superpixels covered by an E-SVM mask into one region. $h(\{s_j|s_j\subset c\})$ is defined as the normalized total

area of the superpixels covered by the template c. The dissimilarity $d(s_1^e, s_2^e)$ in (6) is calculated according to the contextual relationship between the detected contours, that is, $d(s_1^e, s_2^e) = 1$ if there exists any contour across the area covered by s_1^e and s_2^e, otherwise $d(s_1^e, s_2^e) = 0$. (6) can be efficiently solved by using the cutting plane algorithm as introduced in [42]. Intuitively, C is initialized as the whole image, since we have no additional segmentation information at the beginning.

2) Training E-SVMs: The energy $E(w_k, r_k)$ in (4) can be reformulated as the convex energy function of E-SVM as follows:

$$E(w_k, r_k) = \frac{1}{2}||w_k||^2 + \lambda_1 \max(0, 1 - w_k^T f(r_k)) + \lambda_2 \sum_{r_n \in N_E} \max(0, 1 - w_k^T f(r_n)) \quad (7)$$

where N_E indicates the negative examples, and $f(\cdot)$ is the appearance feature of a region, following [1]. λ_1 and λ_2 are two regularization parameters. Thus maximizing $P(W|R)$ in Eqn. (4) is equivalent to minimizing the energy in Eqn. (7), i.e., training the parameters of the E-SVM classifiers by the gradient descent.

We train an E-SVM classifier for each of the selected regions: each selected region is considered as a positive example (exemplar), and a number of patches outside the selected region are cropped as negative examples. In the implementation, we use HOG as the feature for each region. The region selection indicator $\phi(r_j)$ in (4) is determined by the automated saliency detection [43], which aggregates various bottom-up cues and priors to generate spatially coherent yet detail-preserving pixel-accurate and fine-grained saliency, as shown in Fig. 1(a1). For computational efficiency, we only train E-SVMs for high confident foreground regions, i.e., regions containing garment items.

3) Segmentation Propagation: All possible propagations can be searched by sliding window method. However, because we train each E-SVM classifier independently, their responses may not be compatible. Thus the calibration step is performed by fitting a logistic distribution with parameters α_E and β_E on the training set. In this way, the resulting E-SVM response can be computed as

$$S_E(f; w) = \frac{1}{1 + \exp(-\alpha_E(w^T f - \beta_E)} \quad (8)$$

where f is the feature vector of the image patch covered by the sliding window.

B. Co-Labeling

In terms of contextualized clothing co-labeling, each image is described by several coherent regions, and a garment tag is assigned to each region by optimizing a multi-image graphical model. In this paper, the graphical model in (5) is defined as an MRF connecting all the images. We utilize two types of edges on this graph: the interior edges connecting neighboring regions within an image, and the exterior edges connecting regions of different images matched by the propagated segmentation. We show a toy example of the graphical model in Fig. 2.

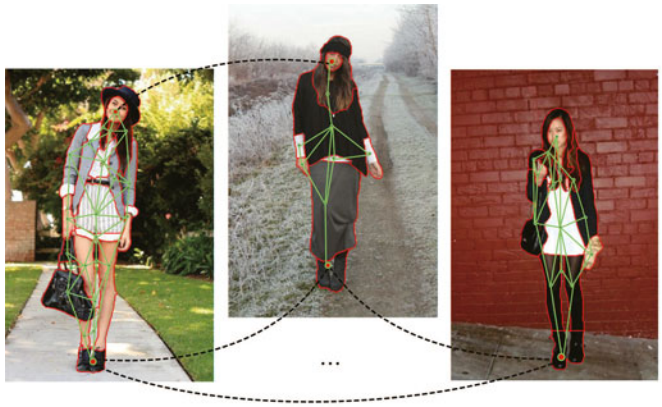


Fig. 2. We perform co-labeling by optimizing a multi-image graphical model, i.e., an MRF connecting all the images in the database. A toy example of the model is illustrated above, where the green solid lines are interior edges between adjacent regions within the same images while the black dashed lines are exterior edges across different images. Note that the connections among different images are determined by the segmentation propagation.

First, the *singleton potential* $P(\ell_k, r_k)$ defined in (5) integrates a region appearance model with the garment item location context. For each garment item, its appearance model is trained as an SVM classifier based on local region appearance.

To classify each region r_k to a specific class of garment, we train the appearance model as a multi-class SVM classifier using one-versus-one decomposition based on local region appearance. Ground-truth label of each region, which is required to train the SVM classifier, is determined by the most frequently occurring pixel label inside the region. Let $S(f(r_k), \ell_k)$ indicate the score of the appearance model, and $f(r_k)$ be a feature vector of 40-bins concatenated by the color and gradient histograms. The singleton potential of assigning label ℓ_k to region r_k can be defined as

$$P(\ell_k, r_k) = \text{sig}(S(f(r_k), \ell_k)) \cdot G_{\ell_k}(X_k) \quad (9)$$

where $\text{sig}(\cdot)$ is the sigmoid function, and X_k denotes the center of region r_k. The location context $G_{\ell_k}(X_k)$ is defined upon the 2-D Gaussian distribution as

$$G_{\ell_k}(X_k) \sim \mathcal{N}(\mu_{\ell_k}, \Sigma_{\ell_k}) \quad (10)$$

where μ_{ℓ_k} and Σ_{ℓ_k} indicate the mean and the covariance of the location of garment item ℓ_k, respectively, which can be estimated by the training set.

There are two types of pairwise potentials considered in constructing the graphical model, that is, interior affinity and exterior affinity respectively.

We define the *interior affinity* $P(\ell_m, \ell_n, r_m, r_n)$ in (5) of two adjacent regions r_m and r_n by considering their appearance compatibility and mutual interactions within an image. The interior affinity can be computed as

$$P(\ell_m, \ell_n, r_m, r_n) = \Phi(\ell_m, \ell_n, r_m, r_n)\Psi(\ell_m, \ell_n). \quad (11)$$

Fig. 3. Some exemplar image pairs in our cross-domain clothing retrieval dataset. For each pair, the image on the left is the online image on the shopping website, and the right one is the photo uploaded by customers, i.e., offline image. These images appear with diverse human poses, illumination, and varying background clutters.

Specifically, the appearance compatibility function $\Phi(\ell_m, \ell_n, r_m, r_n)$ encourages regions with similar appearance to be assigned as the same tag

$$\Phi(\ell_m, \ell_n, r_m, r_n) = \exp\{-\mathbf{1}(\ell_m = \ell_n) d(r_m, r_n)\} \quad (12)$$

where $\mathbf{1}(\cdot)$ is the indicator function, and $d(r_m, r_n)$ is the $\mathcal{X}^2$-distance between the appearance feature of two regions.

The mutual interactions of two different garment items ℓ_m and ℓ_n are modeled by $\Psi(\ell_m, \ell_n)$. $\Psi(\ell_m, \ell_n)$ accumulates the frequency of they appearing as neighbors over all adjacent image patches in the training data. The computation is simple yet effective, since some garments are likely to appear as neighbors in an image, e.g., $coat$ and $pants$, while others are not, e.g., hat and $shoes$.

The *exterior affinity* $Q(\ell_u, \ell_v, r_u, r_v|\mathbf{C})$ of (5) across different images constrains that regions of different images that share similar appearance and locations should have high probability to be assigned as the same garment tag. It can be thus defined as

$$Q(\ell_u, \ell_v, r_u, r_v|\mathbf{C}) = G_{\ell_u}(X_u) G_{\ell_v}(X_v) \Phi(\ell_u, \ell_v, r_u, r_v) \quad (13)$$

where each individual term has been clearly defined in (10) and (12). Finally, the Graph Cuts algorithm is adopted to optimize the multi-image graphical model. The final clothing parsing result can thus be generated according to the outputs from this graphical model.

IV. Application: Clothing Retrieval

In this section, we will show how we facilitate cross-domain clothing retrieval by employing our fine-grained clothing parsing.

A. Data Collection

Notably, current specialized e-commercial websites (e.g., amazon.com, tmall.com) usually allow customers to post their photos with the clothing they have purchased, except exhibiting photos of on sale clothing. Thus, we can collect a large set of online-offline image pairs by crawling the web data. Specifically, the online image indicates the product photos uploaded by merchants and the offline image is the one posted by customers. Initially, we crawled online-offline image pairs from the customer review pages and request several annotators to remove unsuitable images (e.g., heavily occluded or very low resolutions). We have collected about 10 000 online-offline upper-clothing image pairs in high-resolution (about 800 $\times$ 500 on average) from the shopping website, i.e., tmall.com. By extracting from the text tags of images, semantic attribute category (e.g., color) and the specific attribute types for each category (e.g., red, black, white) can be conveniently obtained by parsing these tags. In this dataset, six categories of clothing attributes (i.e., clothing category, clothing color, clothing length, clothing shape, collar shape and sleeve length) and 124 attribute types of all categories are collected. Note that for different attribute category, there are different number of attribute types, i.e., 20 types for clothing category (e.g., T-shirt, Coat), 56 types for color attribute (e.g., black, white), six types for clothing length (e.g., long, short), ten types for clothing shape (e.g., slim, loose), 25 types for collar shape (e.g., round, lapel) and seven types for sleeve length (e.g., long, sleeveless). Some examples of online-offline image pairs are presented in Fig. 3. As can be seen, the online image and its corresponding offline image often appear with different scenarios, various poses, lighting, and background clutters while the fashion images often contain the whole human body and show with high resolution.

Note that our main target is the clothing retrieval task, which aims to match the online images for each query image. It highly relies on the good feature representation for clothing under diverse scenarios. Our large-scale dataset annotated with fine-grained clothing attribute types enables to learn a powerful semantic representation of clothing.

B. Clothing Retrieval Via Fine-Grained Parsing

Given a query image, we first parse it into semantic clothing regions for both online and offline images using the proposed CCP method, and the parsing model is trained based on the proposed dataset. Note that only the upperbody images are included in the retrieval dataset. Thus, only segmentation regions labeled with upperbody related labels are used for describing the image. Among all tags in the dataset, ten tags including cardigan, top, blazer, coat, dress, blouse, jackets, shirt, t-shirt, sweater, are selected. After obtaining the parsing results, all segments of these ten tags are combined into a complete upperbody region. The region masks of upperbody are cropped and fed as the inputs for extracting feature representation, and the background pixels within the masks are set as the mean image values. Benefiting from the parsing results of the photos, the background clutters and body parts can be excluded for feature representation of clothing which enables more accurate clothing localization and high-level information abstraction.

It is well known that deep CNNs have achieved dramatic performances in many areas of computer vision [44],[45], including the attribute prediction [46] and image retrieval [8]. In this work, we utilize the deep features for describing the parsed semantic regions. Note that the semantic attributes included in the clothing retrieval dataset can provide a powerful semantic level representation and help obtain better high-level features for clothing. In this way, we fine-tune the publicly available ImageNet pretrained VGG-16 classification network [27] for attributes prediction. Specifically, we transmit the last fully connected layer into several branches, where each branch corresponds to the classification layer for each attribute category. The neuron number in the output-layer of each branch equals to the number of attribute types of all categories (i.e., 124 types). For each branch, we use the cross-entropy loss as the objective function because each attribute category can have only one activated attribute type. If some attribute categories are missed for some images, the gradients from the corresponding attributes are set as zeros. The loss weight for each branch is set as 1. Following the previous works [4], [47], we use the response of first fully connected layer as feature representation for parsed regions. We then use the ℓ_2 normalization to obtain the final 4096-D features. At last, the Euclidean distance between query and gallery image is calculated and used as the ranking criterion according to the relevance to the query.

V. Experiments

We first introduce the clothing parsing datasets, and present the quantitative results and comparisons. Then the comparison results on clothing retrieval based on clothe parsing are reported.

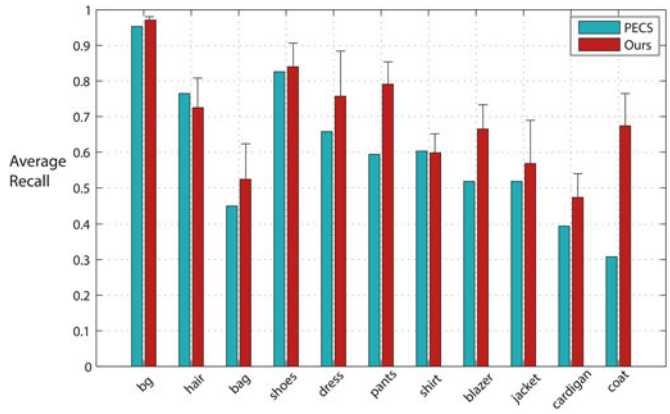


Fig. 4. Average recall of some garment items with high occurrences in Fashionista. The standard deviation measure (vertical bar) is also shown for the average recall of each label.

TABLE I
CLOTHING PARSING RESULTS (%) ON FASHIONISTA AND SYSU-CLOTHES DATASETS

	Fashionista		SYSU-Clothes	
Methods	aPA	mAGR	aPA	mAGR
PECS [2]	89.00	64.37	85.97	51.25
BSC [35]	82.34	33.63	81.61	38.75
STF [49]	68.02	43.62	66.85	40.70
Baseline	77.63	9.03	77.60	15.07
Ours-exterior	89.69	61.26	87.12	61.22
Ours-interior	88.55	61.13	86.75	59.80
Ours-grouping	84.44	47.16	85.43	42.50
Ours-full	**90.29**	**65.52**	**88.23**	**63.89**

We compare our full system (Ours-full) to three previous methods and other versions.

Some qualitative results of clothe parsing and clothing retrieval are exhibited as well.

A. Implementation Details:

We utilize the public gPb contour detector [48] to produce superpixels and contours, and the threshold of the detector is adapted to obtain about 500 superpixels for each image. Contours help define $d(s_1^e, s_2^e)$ in (6) were obtained by setting the threshold to 0.2. For training E-SVMs, we set $\lambda_1 = 0.5$ and $\lambda_2 = 0.01$ in (7) to train E-SVMs. The appearance model in Section III-B is trained by a multi-class SVM using one-against-one decomposition with an Gaussian kernel.

B. Clothing Parsing Datasets

Our clothe parsing framework is extensively evaluated on two datasets: SYSU-Clothes[2] and Fashionista [2]. SYSU-Clothes is a newly constructed dataset by us, which consists of 2098 high-resolution fashion photos with huge human/clothing variations, e.g., in a wide range of styles, accessories, garments, and poses. More than 1000 images of SYSU-Clothes are annotated with

[2][Online]. Available: http://vision.sysu.edu.cn/projects/clothing-co-parsing/

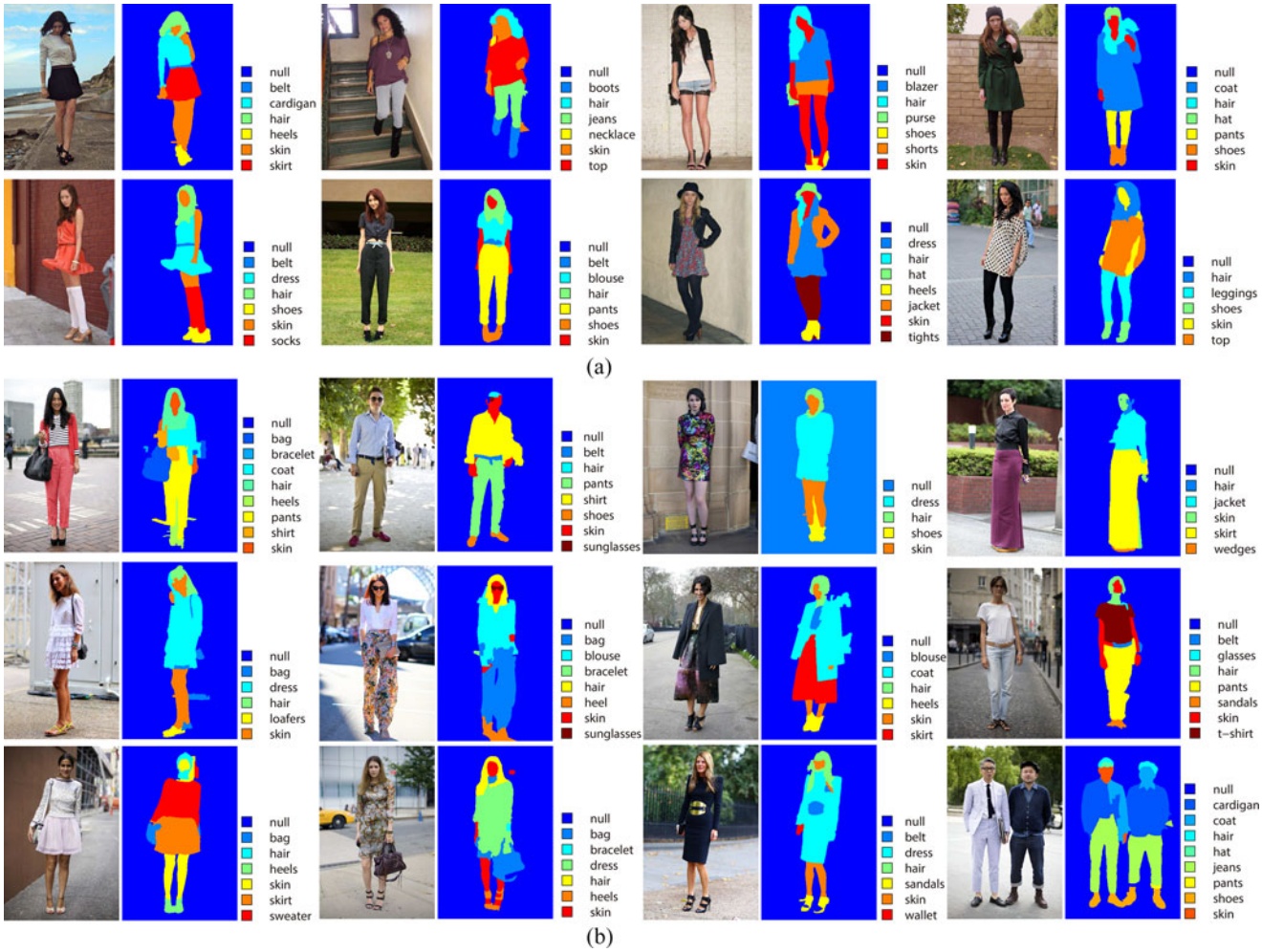


Fig. 5. Some successful parsing results on (a) Fashionista and (b) SYSU-Clothes Better viewed in color.

superpixel-level labeling with totally 57 tags, and the rest of images are only annotated with image-level tags. Some examples are shown in Fig. 5. Fashionista contains 158 235 fashion photos from fashion blogs which are further separated into an annotated subset containing 685 images with superpixel-level ground truth, and an unannotated subset associated with possibly noisy and incomplete tags provided by the bloggers. The annotated subset of Fashionista contains 56 labels, and some garments with high occurrences in the dataset are shown in Fig. 4.

C. Quantitative Evaluation

To evaluate the effectiveness of our framework, we compare our method with three state-of-art methods: 1) PECS [2] which is a fully supervised clothing parsing algorithm that combines pose estimation and clothing segmentation, 2) the BSC [35] for uncovering the label for each image region, and (3) the semantic texton forest (STF) [49], a standard pipeline for semantic labeling. Note that, we can not compare our performance with Dong *et al.* [10] because their results are evaluated on the merged 18 labels instead of using full clothing label set as we used.

The experiment is conducted both on Fashionista and SYSU-Clothes datasets. Following the evaluation protocol in [2], all measurements use ten-fold cross validation, thus nine folds for training as well as for tuning free parameters, and the remaining for testing. The performances are measured by average Pixel Accuracy (aPA) and mean Average Garment Recall (mAGR), as in [2]. As *background* is the most frequent label appearing in the datasets, simply assigning all regions to be *background* achieves 77.63% / 77.60% accuracy, and 9.03% / 15.07% mAGR, on Fashionista and SYSU-Clothes dataset respectively. We treat them as the baseline results.

Table I reports the clothing parsing performance of each method on the Fashionista and SYSU-Clothes datasets. On both datasets, our method achieves much superior performances over the BSC and STF methods, as they did not address the specific clothing knowledge. We also outperform the clothing parsing system PECS on both datasets. As images of SYSU-Clothes include more complex backgrounds and clothing styles, the advantage of our approach is better demonstrated. In fact, the process of iterative image co-segmentation effectively suppresses the image clutters and generates coherent regions, and the co-labeling phase handles better the variants of clothing styles by incorporating rich priors and contexts. In addition, we report the average recall of several frequently occurring garment items in Fashionista dataset in Fig. 4.

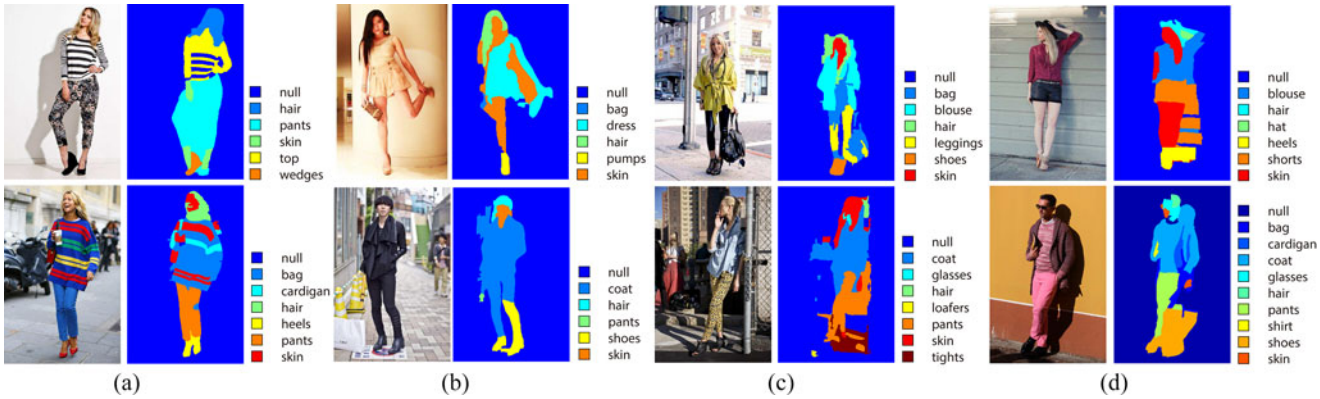


Fig. 6. Some failure cases on Fashionista (first row) and SYSU-Clothes (second row). Best viewed in color.

Evaluation of Components. We also present an empirical analysis to demonstrate the effectiveness of the main components of our system. Ours-exterior and Ours-interior in Table I evaluate the effectiveness of the co-labeling phase by only employing the exterior affinity, and by only using the interior affinity, respectively. Ours-grouping evaluates the performance of superpixel grouping in the co-segmentation phase. Ours-exterior achieves the best result compared to Ours-interior and Ours-grouping due to the importance of mutual interactions between garment items, thus performing co-labeling on a multi-image graphical model benefits the clothing parsing problem.

D. Visualization

Fig. 5 shows some successful clothing parsing results for exemplary images from both Fashionista and SYSU-Clothes. It can be observed that our framework can be able to parse clothing accurately even in some challenging illumination and complex background clutters (r1c2,[3] r4c2). Moreover, our framework can also successfully parse some small garments such as *belt* (r1c1, r2c1, r2c2, r3c2), *purse* (r1c3), *hat* (r1c4, r2c3), and *sunglasses* (r4c2). For reasonably ambiguous clothing patterns such as dotted t-shirt or colorful dress, our framework could give satisfying results (r2c4, r5c2). In addition, the proposed method could parse the images with several persons in a single image simultaneously (r5c5).

Some failure cases are illustrated in Fig. 6. Our co-parsing framework may lead to wrong results under following scenarios: 1) ambiguous patterns exist within a clogging garment item; 2) different clothing garment items share similar appearance; 3) background is extremely disordered; 4) illumination condition is poor.

E. Efficiency

We conduct all the experiments on an Intel Dual-Core E6500 (2.93 GHz) CPU and 8GB RAM PC. The run-time complexity of the co-segmentation phase scales linearly with the number of iterations, and each iteration costs about 10 s/image. The co-labeling phase costs less than 1 min to assign labels to a database of 70 images, which is very effective due to the consistent regions obtained from the co-segmentation phase. And the Graph Cuts algorithm converges in 3–4 iterations in our experiment.

[3]We use "r1c1" to denote the image in row 1, column 1.

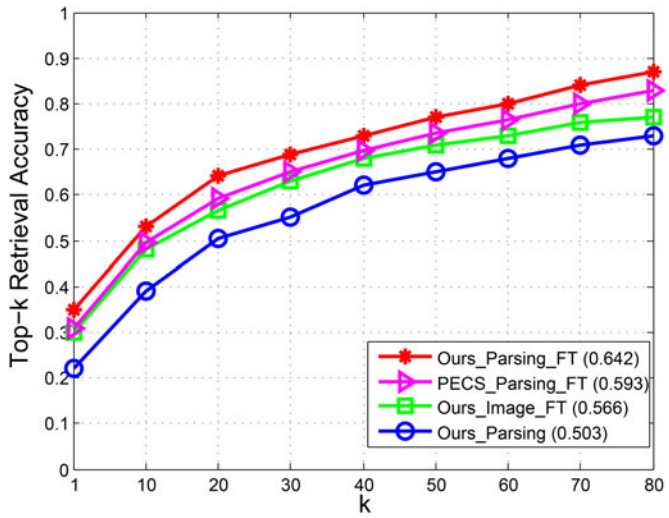


Fig. 7. Top-k retrieval accuracy on 5000 retrieval gallery images. We test nine different numbers of k range from 1 to 80. The number in the parentheses is the top-20 retrieval accuracy.

F. Evaluation on Clothing Retrieval

For training the domain-specific network for clothing attribute prediction, about 5000 online-offline image pairs are randomly selected for network training. During the training process, we augment the training images with the horizontal reflections, and the network is trained and tested based on Caffe [50] on a single NIVIDIA Tesla K40c. The network is trained using stochastic gradient descent with a batch size of 20 images, momentum of 0.9, and weight decay of 0.0005. The learning rate for classification layers is initialized at 0.001 and divided by 10 after 30 epochs, and the learning rate of other layers initialized by VGG network is set as 0.0001. We train the network for roughly 90 epochs. For testing, we used 5000 online-offline image pairs, and the offline images from customers are treated as queries and online images are used as the retrieval gallery. On average, the attribute-aware feature extraction process costs about 0.01 s per image. Given a parsed semantic region of the query image, it

Fig. 8. Top-6 retrieval result of our method. The images in first column are the queries, and the retrieved images with red boxes are the same clothing images. Best viewed in original PDF file.

costs about 0.5 s for feature extraction and clothing retrieval in our experiment.

To evaluate the effectiveness of the usage of parsing results for clothing retrieval, we compare our full version ("Ours_Parsing_FT") with the version ("Ours_Image_FT") that extracts features for the whole image instead of parsed semantic regions. To further analyze the retrieval performance of deep features, "Ours_Parsing" reports the results by directly using the pre-trained VGG network to extract the deep features for clothing regions. The retrieval performance of using parsing results by PECS [2] is also reported as "PECS_Parsing_FT" to enable further comparison with our CCP method. Following the previous retrieval method [4], we used the top-k retrieval accuracy in which we denote a hit if we find the exact same clothing in the top k results otherwise a miss.

We give full detailed top-k retrieval accuracy results for different methods in Fig. 7. The k is varied as it is an important indicator for a real retrieval system. The top-20 retrieval accuracy of each model is listed in the corresponding parentheses. Our method ("Ours_Parsing_FT") achieves superior retrieval performance over the version ("PECS_Parsing_FT") based on the baseline [2], i.e., 64.2% versus 59.3% on top-20 retrieval accuracy. Compared to "Ours_Parsing," the finetuned network by using attribute annotations can significantly improve the retrieval performance by 13.9%. This attests the effectiveness of attributes for learning powerful semantic features for clothing

retrieval. By comparing with the baseline ("Ours_Image_FT") that uses the whole image for extracting clothing features, our parsing-based clothing retrieval method greatly improve the performance by 7.6%. It verifies well that our CCP method can be treated as a effective pre-processing step for better clothing retrieval performance. Some exemplar retrieval results are illustrated in Fig. 8.

VI. Conclusion

This paper has presented a novel system for jointly parsing a batch of clothing images given the image-level clothing tags. Our framework consists of two phases of inference: image co-segmentation and region co-labeling. The large high-resolution street fashion photos dataset annotated with pixel-wise labeling and fine-grained clothing tags is made available to the public for promoting further academic research on clothing analysis. The experiments demonstrate that our framework is effective and applicable compared with the existing methods. In addition, the parsing-based clothing retrieval pipeline has also been proposed to utilize the clothing parsing results for clothing retrieval. The large cross-domain image retrieval dataset with 10 000 online-offline image pairs has been proposed. The significant improvement on retrieval performance over the baselines further verifies well the effectiveness of our parsing framework. In future work, we plan to utilize our framework for the real-world clothing-related applications, such as virtual outfit try-on system and clothing attribute prediction. In addition, we will extend our framework on generic image segmentation tasks and incorporate deep learning architecture into our framework to further boost the performances. In addition, the parallel implementation would be studied to adapt the large scale applications.

References

[1] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of exemplar-SVMs for object detection and beyond," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 89–96.

[2] K. Yamaguchi, H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 3570–3577.

[3] M. Yuan, I. Khan, F. Farbiz, S. Yao, A. Niswar, and M.-H. Foo, "A mixed reality virtual clothes try-on system," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1958–1968, Dec. 2013.

[4] J. Huang, R. S. Feris, Q. Chen, and S. Yan, "Cross-domain image retrieval with a dual attribute-aware ranking network," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1062–1070.

[5] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun, "Neuroaesthetics in fashion: Modeling the perception of fashionability," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 869–877.

[6] V. Jagadeesh, R. Piramuthu, A. Bhardwaj, W. Di, and N. Sundaresan, "Large scale visual recommendations from street fashion images," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 1925–1934.

[7] Y. Kalantidis, L. Kennedy, and L.-J. Li, "Getting the look: Clothing recognition and segmentation for automatic product suggestions in everyday photos," in *Proc. 3rd ACM Conf. Int. Conf. Multimedia Retrieval*, 2013, pp. 105–112.

[8] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang, "Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4766–4779, Dec. 2015.

[9] H. Chen, A. Gallagher, and B. Girod, "Describing clothing by semantic attributes," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 609–623.

[10] J. Dong, Q. Chen, W. Xia, Z. Huang, and S. Yan, "A deformable mixture parsing model with parselets," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3408–3415.

[11] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Retrieving similar styles to parse clothing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 5, pp. 1028–1040, May 2015.

[12] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun, "A high performance CRF model for clothes parsing," in *Proc. Asian Conf. Comput. Vis.*, 2015, pp. 64–81.

[13] G. Kim, E. Xing, L. Fei-Fei, and T. Kanade, "Distributed cosegmentation via submodular optimization on anisotropic diffusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 169–176.

[14] A. Gallagher and T. Chen, "Clothing cosegmentation for recognizing people," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–8.

[15] L. Lin, T. Wu, J. Porway, and Z. Xu, "A stochastic graph grammar for compositional object representation and recognition," *Pattern Recog.*, vol. 42, no. 7, pp. 1297–1307, 2009.

[16] X. Liu, L. Lin, S. Yan, H. Jin, and W. Tao, "Integrating spatio-temporal context with multiview representation for object recognition in visual surveillance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 4, pp. 393–407, Apr. 2011.

[17] L. Lin *et al.*, "Object categorization with sketch representation and generalized samples," *Pattern Recog.*, vol. 45, no. 10, pp. 3648–3660, 2012.

[18] D. Kuettel, M. Guillaumin, and V. Ferrari, "Segmentation propagation in imagenet," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 459–473.

[19] A. Vezhnevets, V. Ferrari, and J. Buhmann, "Weakly supervised semantic segmentation with a multi-image model," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 643–650.

[20] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Recog. Image Anal.*, vol. 96, no. 1, pp. 1–27, Nov. 2001.

[21] A. Borras, F. Tous, J. Lladós, and M. Vanrell, "High-level clothes description based on colour-texture and structural features," in *Proc. 1st Iberian Conf. Pattern Recog. Image Anal.*, 2003, pp. 108–116.

[22] H. Chen, Z. J. Xu, Z. Q. Liu, and S. C. Zhu, "Composite templates for cloth modeling and sketching," in *Proc. IEEE Comput. Sci. Conf. Comput. Vis. Pattern Recog.*, Jun. 2006, vol. 1, pp. 943–950.

[23] X. Liang *et al.*, "Deep human parsing with active template regression," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 2, pp. 2402–2414, Dec. 2015.

[24] S. Liu *et al.*, "Matching-CNN meets KNN: Quasi-parametric human parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 1419–1427.

[25] X. Wang and T. Zhang, "Clothes search in consumer photos via color matching and attribute learning," in *Proc. ACM Int. Conf. Multimedia*, 2011, pp. 1353–1356.

[26] X. Wang, T. Zhang, D. R. Tretter, and Q. Lin, "Personal clothing retrieval on photo collections by color and attributes," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 2035–2045, Dec. 2013.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556.

[28] B. Hasan and D. Hogg, "Segmentation using deformable spatial priors with application to clothing," in *Proc. Brit. Mach. Vis. Conf.*, 2010, pp. 83.1–83.11.

[29] Y. Bo and C. C. Fowlkes, "Shape-based pedestrian parsing," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, Jun. 2011, pp. 2265–2272.

[30] N. Wang and H. Ai, "Who blocks who: Simultaneous clothing segmentation for grouping images," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1535–1542.

[31] P. Luo, X. Wang, and X. Tang, "Hierarchical face parsing via deep learning," in *Proc. Comput. Vis. Pattern Recog.*, 2012, pp. 2480–2487.

[32] P. Luo, X. Wang, and X. Tang, "Pedestrian parsing via deep decompositional network," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2648–2655.

[33] S. Liu *et al.*, "Fashion parsing with video context," *IEEE Trans. Multimedia*, vol. 17, no. 8, pp. 1347–1358, Aug. 2015.

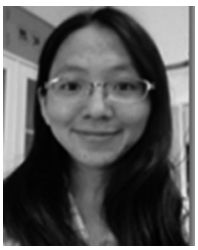
[34] S. Liu *et al.*, "Fashion parsing with weak color-category labels," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 253–265, Jan. 2014.

[35] X. Liu *et al.*, "Label to region by bi-layer sparsity priors," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 115–124.

[36] P. Luo, X. Wang, L. Lin, and X. Tang, "Joint semantic segmentation by searching for compatible-competitive references," in *Proc. ACM Int. Conf. Multimedia*, 2012, pp. 777–780.

[37] L. Lin, X. Liu, and S.-C. Zhu, "Layered graph matching with composite cluster sampling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1426–1442, Aug. 2010.

[38] B. Leibe, A. Leonardis, and B. Schiele, "Combined object categorization and segmentation with an implicit shape model," in *Proc. Workshop Statist. Learn. Comput. Vis., Eur. Conf. Comput. Vis.*, 2004, pp. 17–32.
[39] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *Proc. 10th IEEE Int. Conf. Comput. Vis.*, Oct. 2005, vol. 2, pp. 1800–1807.
[40] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 584–599.
[41] W. Yang, P. Luo, and L. Lin, "Clothing co-parsing by joint image segmentation and labeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 3182–3189.
[42] S. Nowozin and S. Jegelka, "Solution stability in linear programming relaxations: Graph partitioning and unsupervised learning," in *Proc. 26th Annu. Int. Conf. Int. Conf. Mach. Learn.*, 2009, pp. 769–776.
[43] K. Wang, L. Lin, J. Lu, C. Li, and K. Shi, "Pisa: Pixelwise image saliency by aggregating complementary appearance contrast measures with edge-preserving coherence," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 1057–7149, Oct. 2015.
[44] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 1–9.
[45] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in *Proc. Int. Conf. Representation Learn.*, 2014
[46] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, "PANDA: Pose aligned networks for deep attribute modeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 1637–1644.
[47] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *Proc. Int. Conf. Representation Learn.*, 2013.
[48] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011.
[49] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–8.
[50] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.



Xiaodan Liang is currently working toward the Ph.D. degree at Sun Yat-Sen University, Guangzhou, China.

She is currently a Research Intern with the National University of Singapore, Singapore. Her research interests mainly include semantic segmentation, object/action recognition, and medical image analysis.

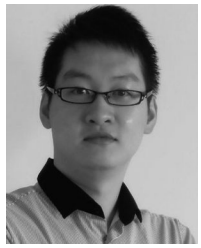
Liang Lin received the Ph.D. degree from the Beijing Institute of Technology (BIT), Beijing, China, in 2008.

He is a Professor of computer science with Sun Yat-Sen University (SYSU), Guangzhou, China. He was a Postdoctoral Research Fellow with the Center for Vision, Cognition, Learning, and Art, University of California at Los Angeles, Los Angeles, CA, USA. He has authored or coauthored more than 80 papers in academic journals and conferences. His research interests include new models, algorithms, and systems for intelligent processing and understanding of visual data such as images and videos.

Prof. Lin has served as an Associate Editor of *Neurocomputing* and *The Visual Computer*. He was supported by several promotive programs or funds for his works, such as the Program for New Century Excellent Talents of Ministry of Education (China) in 2012 and the Guangdong NSFs for Distinguished Young Scholars in 2013. He was the recipient of the Best Paper Runners-Up Award in ACM NPAR 2010, the Google Faculty Award in 2012, and the Best Student Paper Award in IEEE ICME 2014.

Wei Yang received the B.S. degree in software engineering and the M.S. degree in computer software and theory from Sun Yat-Sen University, Guangzhou, China, in 2011 and 2014, respectively, and is currently working toward the Ph.D. degree in electronic engineering from the Chinese University of Hong Kong, Hong Kong, China.

His research interest includes computer vision and deep learning.

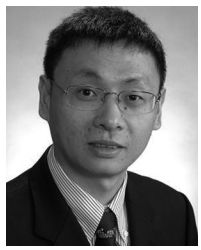


Ping Luo received the Ph.D. degree in information engineering from the Chinese University of Hong Kong, Hong Kong, in 2014.

He is currently a Postdoctoral Researcher. His research interests include computer vision, pattern recognition, and machine learning, focusing on face analysis, and large-scale object recognition and detection.

Junshi Huang received the B.S. and M.S. degrees from the Beijing Institute of Technology, Beijing, China, in 2009 and 2012, respectively, and is currently working toward the Ph.D. degree in electrical and computer engineering at the National University of Singapore, Singapore.

His research interests include object detection and object retrieval.



Shuicheng Yan is currently an Associate Professor with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, and the Founding Lead of the Learning and Vision Research Group. He has authored or coauthored nearly 400 technical papers over a wide range of research topics, with Google Scholar citation more than 12,000 times. His research interests include machine learning, computer vision, and multimedia.

Prof. Yan is an ISI highly-cited Researcher (2014), and an IAPR Fellow (2014). He has been serving as an Associate Editor of the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, *Computer Vision and Image Understanding*, and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. He was the recipient of the Best Paper Award from ACM MM'2013 (Best Paper and Best Student Paper), ACM MM'2012 (Best Demo), PCM'2011, ACM MM'2010, ICME'2010, and ICIMCS'2009, the runner-up prize of ILSVRC'2013, the winner prizes of the classification task in PASCAL VOC 2010–2012, the winner prize of the segmentation task in PASCAL VOC 2012, the honorable mention prize of the detection task in PASCAL VOC 2010, 2010 IEEE Transactions on Circuits and Systems for Video Technology Best Associate Editor Award, 2010 Young Faculty Research Award, 2011 Singapore Young Scientist Award, and 2012 NUS Young Researcher Award.