Clothing Co-Parsing by Joint Image Segmentation and Labeling

Wei Yang[1], Ping Luo[2], Liang Lin[1,3*]

[1]Sun Yat-sen University, Guangzhou, China

[2]Department of Information Engineering, The Chinese University of Hong Kong

[3]SYSU-CMU Shunde International Joint Research Institute, Shunde, China

{platero.yang, pluo.lhi}@gmail.com, linliang@ieee.org

Abstract

This paper aims at developing an integrated system of clothing co-parsing, in order to jointly parse a set of clothing images (unsegmented but annotated with tags) into semantic configurations. We propose a data-driven framework consisting of two phases of inference. The first phase, referred as "image co-segmentation", iterates to extract consistent regions on images and jointly refines the regions over all images by employing the exemplar-SVM (E-SVM) technique [23]. In the second phase (i.e. "region co-labeling"), we construct a multi-image graphical model by taking the segmented regions as vertices, and incorporate several contexts of clothing configuration (e.g., item location 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 [30], we construct a dataset called CCP 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 CCP datasets, respectively, which are superior compared with state-of-the-art methods.

1. Introduction

Clothing recognition and retrieval have huge potentials in internet-based e-commerce, as the revenue of online clothing sale keeps highly increasing every year. In computer vision, several interesting works [30, 5, 17, 16] have been proposed on this task and showed promising results.

On one hand, pixelwise labeling of clothing items within images is one of the key resources for the above researches, but it often costs expensively and processes inefficiently. On the other hand, it is feasible to acquire image-level clothing tags based on rich online user data. Therefore, an interesting problem arises, which is the focus of this paper: How to jointly segment the clothing images into regions of clothes and simultaneously transfer semantic tags at image level to these regions.

The key contribution of this paper is an engineered and applicable system[1] to jointly parse a batch of clothing images and produce accurate pixelwise annotation of clothing items. We consider the following challenges to build such a system:

- The appearances of clothes and garment items are often diverse with different styles and textures, compared with other common objects. It is usually hard to segment and recognize clothes via only bottom-up image features.

- The variations of human poses and self-occlusions are non-trivial issues for clothing recognition, although the clothing images can be in clear resolution and nearly frontal view.

- The number of fine-grained clothes categories is very large, *e.g.*, more than 50 categories in the Fashionista dataset [30]; the categories are relatively fewer in existing co-segmentation systems [9, 7].

To address the above issues, we develop the system consisting of two sequential phases of inference over a set of clothing images, *i.e.* image co-segmentation for extracting distinguishable clothes regions, and region co-labeling for recognizing various garment items, as illustrated in Figure 1. Furthermore, we exploit contexts of clothing configuration, *e.g.*, spatial locations and mutual relations of clothes items, inspired by the successes of object/scene context modeling [14, 19, 12].

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

(a) Phase I: Co-Segmentation
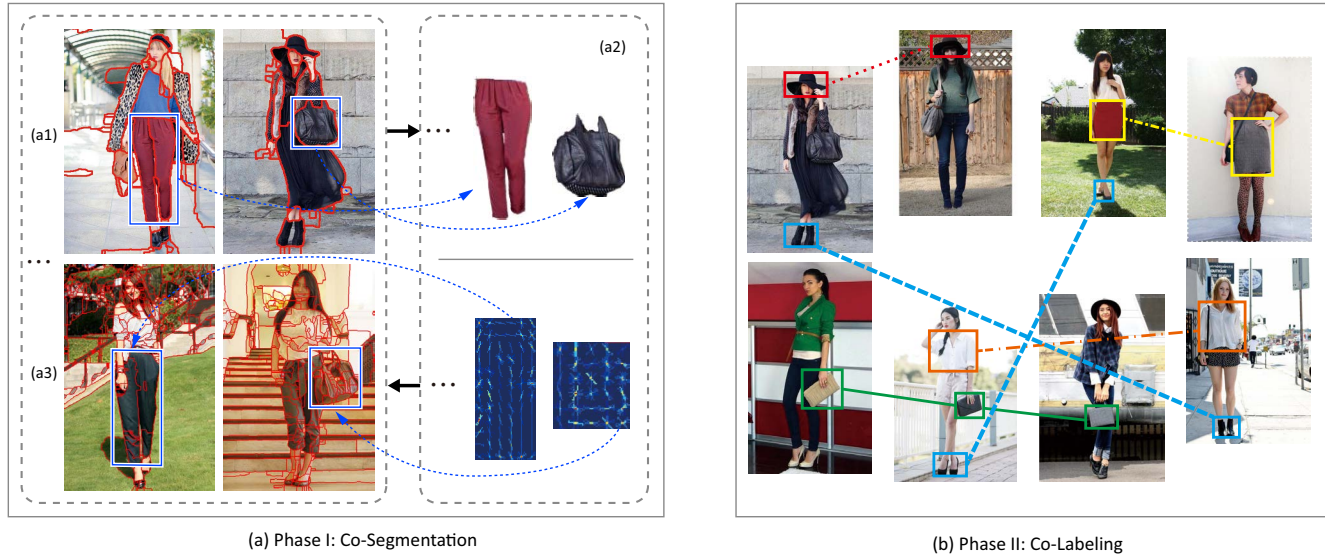
(b) Phase II: Co-Labeling

Figure 1. Illustration of the proposed clothing 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, (a2) selecting confident foreground regions 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).

In the phase of image co-segmentation, the algorithm iteratively refines the regions grouped over all images by employing the exemplar-SVM (E-SVM) technique [23]. First, we extract superpixels and group them into regions for each image, where most regions are often cluttered and meaningless due to the diversity of clothing appearances and human variations, as shown in Figure 1 (a1). Nevertheless, some coherent regions (in Figure 1 (a2)) can be still selected that satisfy certain criteria (*e.g*., size and location). Then we train a number of E-SVM classifiers for the selected regions using the HOG feature, *i.e*., one classifier for one region, and produce a set of region-based detectors, as shown in Figure 1 (a3), which are applied as 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. [10] also proposed to propagate segmentations through HOG-based matching.

Given the segmented regions of all images, it is very difficult to recognize them by only adopting supervised learning due to the large number of fine-grained categories and the large intra-class variance. In contrast, we perform the second phase of co-labeling in a data-driven manner. We construct a multi-image graphical model by taking the regions as vertices of graph, inspired by [27]. In our mod-el, we link adjacent regions within each image as well as regions across different images, which shares similar appearance and latent semantic tags. Thus we can borrow s-tatistical strength from similar regions in different images and assign labels jointly, as Figure 1 (b) illustrates. The optimization of co-labeling is solved by the efficient Graph Cuts algorithm [4] that incorporates several constraints defined upon the clothing contexts.

Moreover, a new database with groundtruths is proposed for evaluating clothing co-parsing, including more realistic and general challenges, *e.g*. disordered backgrounds and multiform human poses, compared with the existing clothes datasets [3, 6, 30]. We demonstrate promising performances and applicable potentials of our system in the experiments.

1.1. Related Work

In literature, existing efforts on clothing/human segmentation and recognition mainly focused on constructing expressive models to address various clothing styles and appearances [6, 8, 2, 28, 21, 30, 22]. One classic work [6] proposed a composite And-Or graph template for modeling and parsing clothing configurations. Later works studied on blocking models to segment clothes for highly occluded group images [28], or deformable spatial priors modeling for improving performance of clothing segmentation [8]. Recent approaches incorporated shape-based human model [2], or pose estimation and supervised region labeling [30], and achieved impressive results. Despite acknowl-

edged successes, these works have not yet been extended to the problem of clothing co-parsing, and they often require much labeling workload.

Clothing co-parsing is also highly related to image/object co-labeling, where a batch of input images containing similar objects are processed jointly [18, 20, 13]. For example, unsupervised shape guided approaches were adopted in [11] to achieve single object category co-labeling. Winn et. al. [29] incoporated 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. [18] proposed to propagate labels over scene images using a bi-layer sparse coding formulation. Similar ideas were also explored in [15]. These methods, however, are often limited by expensive annotations. In addition, they extracted image correspondences upon the pixels (or superpixels), which are not discriminative for the clothing parsing problem.

The rest of this paper is organized as follows. We first introduce the probabilistic formulation of our framework in Section 2, and then discuss the implementation of the two phases in Section 3. The experiments and comparisons are presented in Section 4, and finally comes the conclusion in Section 5.

2. Probabilistic Formulation

We formulate the task of clothing co-parsing 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 will be further grouped into several coherent regions under the guidance of segmentation propagation. Each image I is associated with four additional variables:

a) the regions $\{r_k\}_{k=1}^K$, each of which is consisted of a set of superpixels;

b) the garment label for each region: $\ell_k \in T, k = 1, ..., K$;

c) the E-SVM weights w_k trained for each selected region;

d) 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 Figure 1 (a).

Let $\mathbf{R} = \{R_i = \{r_{ik}\}\}$, $\mathbf{L} = \{L_i = \{\ell_{ik}\}\}$, $\mathbf{W} = \{W_i = \{w_{ik}\}\}$ and $\mathbf{C} = \{C_i\}$. We optimize the parameters 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 as

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

The optimization of Eq. (2) includes two phases: (I) clothing image co-segmentation and (II) co-labeling.

In *phase* (I), we obtain the optimal regions by maximizing $P(R|C, I)$ in Eq. (2). We introduce the superpixel grouping indicator $o_j \in \{1, \cdots, K\}$, which indicates to which of the K regions the superpixel s_j belongs. Then each region can be denoted as a set of superpixels, as $r_k = \{s_j|o_j = k\}$. Given the current segmentation propagation C, $P(R|C, I)$ can be defined 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 belongs to a region, where $d(s_j, o_j)$ evaluates the spatial distance 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, we then select several coherent regions 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, we define $P(C_i|\mathbf{W}, I_i)$ in Eq. (2) based on the responses of E-SVM classifiers. This probability 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 problem as a multi-image graphical model,

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

where $P(\ell_{ik}, r_{ik})$ represents 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 $\mathbf{C}$. Details are discussed in Section 3.2.

3. Clothing Co-Parsing

In this section, we describe the two phases of clothing co-parsing, including their implementation details. The overall procesure is outlined in Algorithm 1.

3.1. 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 between three steps:

i. Superpixel Grouping: The MRF model defined in Eq. (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.

To automatically determine the number of regions, we replace the superpixel indicator o_j by a set of *binary* variables o^e defined on the edges between 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 region, otherwise $o^e = 0$. We also introduce a binary variable o^c with $o^c = 1$ indicating all the superpixels covered by the mask of the segmentation propagation c belong to the same region, otherwise $o^c = 0$. Then maximizing Eq. (3) is equivalent to 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 $d(s_1^e, s_2^e)$ is the dissimilarity between two superpixels, and $h(\cdot)$ measures the consistence of grouping all superpixels covered by an E-SVM mask into one region. Eq. (6) can be efficiently solved via the cutting plane algorithm as introduced in [24].

The dissimilarity $d(s_1^e, s_2^e)$ in Eq. (6) is defined together with the detected contours: $d(s_1^e, s_2^e) = 1$ if there exists any contour across the area covered by s_1^e and s_2^e, otherwise

$d(s_1^e, s_2^e) = 0$. $h(\{s_j | s_j \subset c\})$ is defined as the normalized total area of the superpixels covered by the template c.

ii. Training E-SVMs: The energy $E(w_k, r_k)$ introduced in Eq. (4) is the convex energy function of E-SVM as follows,

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

N_E denotes the negative examples, and $f(\cdot)$ is the feature of a region, following [23]. λ_1 and λ_2 are two regularization parameters. Thus maximizing $P(W|R)$ is equivalent to minimizing the energy in Eq. (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 croped as negative examples. In the implementation, we use HOG as the feature for each region. The region selection indicator $\phi(r_j)$ in Eq. (4) is determined by the automated saliency detection [25]. For computational efficiency, we only train E-SVMs for high confident foreground regions, *i.e.* regions containing garment items.

iii. Segmentation Propagation: We search for possible propagations by sliding window method. However, as each E-SVM is trained independently, their responses may not be compatible. We thus perform the calibration by fitting a logistic distribution with parameters α_E and β_E on the training set. Then the E-SVM response can be defined 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.

3.2. Contextualized Co-Labeling

In this phase, each image is represented by a set of coherent regions, and we assign a garment tag to each region by optimizing a multi-image graphical model , *i.e.* an MRF connecting all the images in the database, which is defined in Eq. (5). We define two types of edges on the graph: the interior edges connecting neighboring regions within an image, and the exterior edges connecting regions of different images matched by the propagated segmentation. A toy example of the graphical model is showed in Figure 2. Specifically, two types of clothing contexts are exploited providing informative constraints during inference.

The **singleton potential** $P(\ell_k, r_k)$ defined in Eq. (5) incorporates a region appearance model with the garment item location context.

For each type of garment, we train the appearance model as an SVM classifier based on local region appearance. Let $S(f(r_k), \ell_k)$ denote the score of the appearance model, and $f(r_k)$ a feature vector of 40-bins concatenated by the
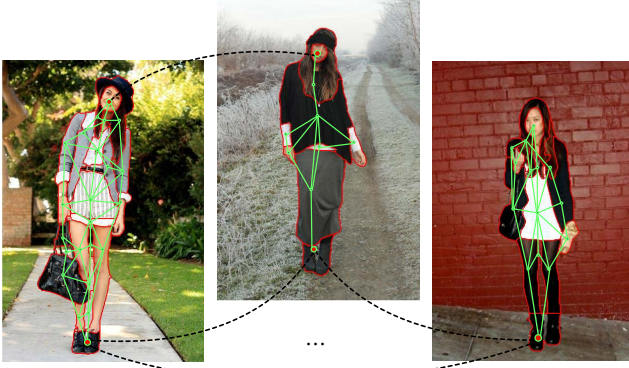
Figure 2. We perform co-labeling by optimizing a multi-image graphical model, *i.e.* an MRF connecting all the images in the database. A toy example of the model is illustrated above, where the green solid lines are interior edges between adjacent regions within the same images while the black dashed lines are exterior edges across different images. Note that the connections among different images are determined by the segmentation propagation.

color and gradient histograms. We define the potential of assigning label ℓ_k to region r_k as,

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

where $\text{sig}(\cdot)$ indicates the sigmoid function, and X_k 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} represent the mean and the covariance of the location of garment item ℓ_k, respectively, which can be estimated over the training set.

The **interior affinity** $P(\ell_m, \ell_n, r_m, r_n)$ in Eq. (5) of two adjacent regions r_m and r_n is defined on two terms within an image: their appearance compatibility and mutual interactions, as

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

The appearance compatibility function $\Phi(\ell_m, \ell_n, r_m, r_n)$ encourages regions with similar appearance to have 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.

$\Psi(\ell_m, \ell_n)$ models the mutual interactions of two different garment items ℓ_m and ℓ_n. This term is simple but 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*. In practice, we calculate $\Psi(\ell_m, \ell_n)$ by accumulating the frequency of they appearing as neighbors over all adjacent image patches in the training data.

The **exterior affinity** $Q(\ell_u, \ell_v, r_u, r_v|\mathbf{C})$ of Eq. (5) across different images constrains that regions in different images sharing similar appearance and locations should have high probability to have the same garment tag. We thus have,

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

in which the terms were clearly defined in Eq.(10) and Eq.(12). Finally, we adopt the Graph Cuts to optimize the multi-image graphical model.

Algorithm 1 The Sketch of Clothing 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
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 Eq. (3);

  2 Train E-SVM parameters for each selected region by minimizing the energy in Eq. (7).

  3 Propagate segmentations across images by detections from the trained E-SVM classifiers by Eq. (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 Eq. (5) by Graph Cuts.

---

4. Experiments

We first introduce the clothing parsing datasets, and present the quantitative results and comparisons. Some qualitative results are exhibited as well.

Parameter settings: We use gPb contour detector [1] to obtain 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 Eq. (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 Eq. (7) to train E-SVMs. The appearance model in Sec. 3.2 is trained by a multi-class SVM using one-against-one decomposition with an Gaussian kernel.

4.1. Clothing Image Datasets

We evaluate our framework on two datasets: Clothing Co-Parsing[2] (CCP) and Fashionista [30]. CCP is created by us, consisting of $2,098$ high-resolution fashion photos with huge human/clothing variations, which are in a wide range of styles, accessories, garments, and poses. More than 1000 images of CCP are with superpixel-level annotations with totally 57 tags, and the rest of images are annotated with image-level tags. All annotations are produced by a professional team. Some examples of CCP are shown in Figure 4. 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 Figure 3.

4.2. Quantitative Evaluation

To evaluate the effectiveness of our framework, we compare our method with three state-of-arts: (1) PECS [30] which is a fully supervised clothing parsing algorithm that combines <u>p</u>ose <u>e</u>stimation and <u>c</u>lothing <u>s</u>egmentation, (2) the <u>b</u>i-layer <u>s</u>parse <u>c</u>oding (BSC) [18] for uncovering the label for each image region, and (3) the <u>s</u>emantic <u>t</u>exton <u>f</u>orest (STF) [26], a standard pipeline for semantic labeling.

The experiment is conducted both on Fashionista and CCP datasets. Following the protocol in [30], all measurements use 10-fold cross validation, thus 9 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 [30]. 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 the Fashionista and the CCP dataset respectively. We treat them as the baseline results.

Table 1 reports the clothing parsing performance of each method on the Fashionista and CCP datasets. On both datasets, our method achieves much superior performances over the BSC and STF methods, as they did not address the specific clothing knowledges. We also outperform the state-of-the-art clothing parsing system PECS on both datasets. As images of the CCP database 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 aver-

Methods	Fashionista		CCP	
	aPA	mAGR	aPA	mAGR
Ours-full	**90.29**	**65.52**	**88.23**	**63.89**
PECS [30]	89.00	64.37	85.97	51.25
BSC [18]	82.34	33.63	81.61	38.75
STF [26]	68.02	43.62	66.85	40.70
Ours-1	89.69	61.26	87.12	61.22
Ours-2	88.55	61.13	86.75	59.80
Ours-3	84.44	47.16	85.43	42.50
Baseline	77.63	9.03	77.60	15.07

Table 1. Clothing parsing results (%) on the Fashionista and the CCP dataset. As *background* is the most frequent label appears in the datasets, assigning all regions to be *background* is adopted our baseline comparison. We compare our full system (Ours-full) to three state-of-the-art methods including PECS [30], BSC [18], and STF [26]. We also present an empirical analysis to evaluate the effectiveness of the main components of our system. Ours-1 and Ours-2 evaluate the effectiveness of the co-labeling phase by only employing the exterior affinity, and by only using the interior affinity, respectively. Ours-3 evaluates the performance of superpixel grouping in the co-segmentation phase.
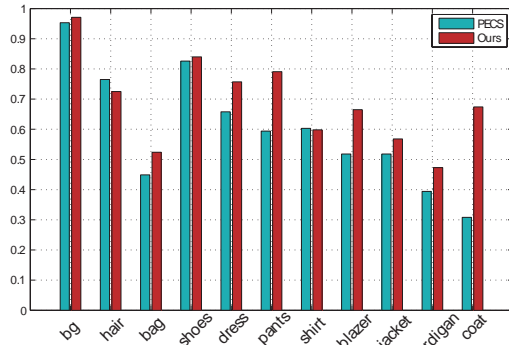

Figure 3. Average recall of some garment items with high occurrences in Fashionista.

age recall of several frequently occurring garment items in Fashionista dataset in Figure 3.

Evaluation of Components. We also present an empirical analysis to demonstrate the effectiveness of the main components of our system. Ours-1 and Ours-2 in Table 1 evaluate the effectiveness of the co-labeling phase by only employing the exterior affinity, and by only using the interior affinity, respectively. Ours-3 evaluates the performance of superpixel grouping in the co-segmentation phase. Ours-1 achieves the best result compared to Ours-2 and Ours-3 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.

4.3. Qualitative Evaluation

Figure 4 illustrates some successful parsing results for exemplary images from both Fashionista and CCP. Our method is able to parse clothes accurately even in some challenging illumination and complex background condi-
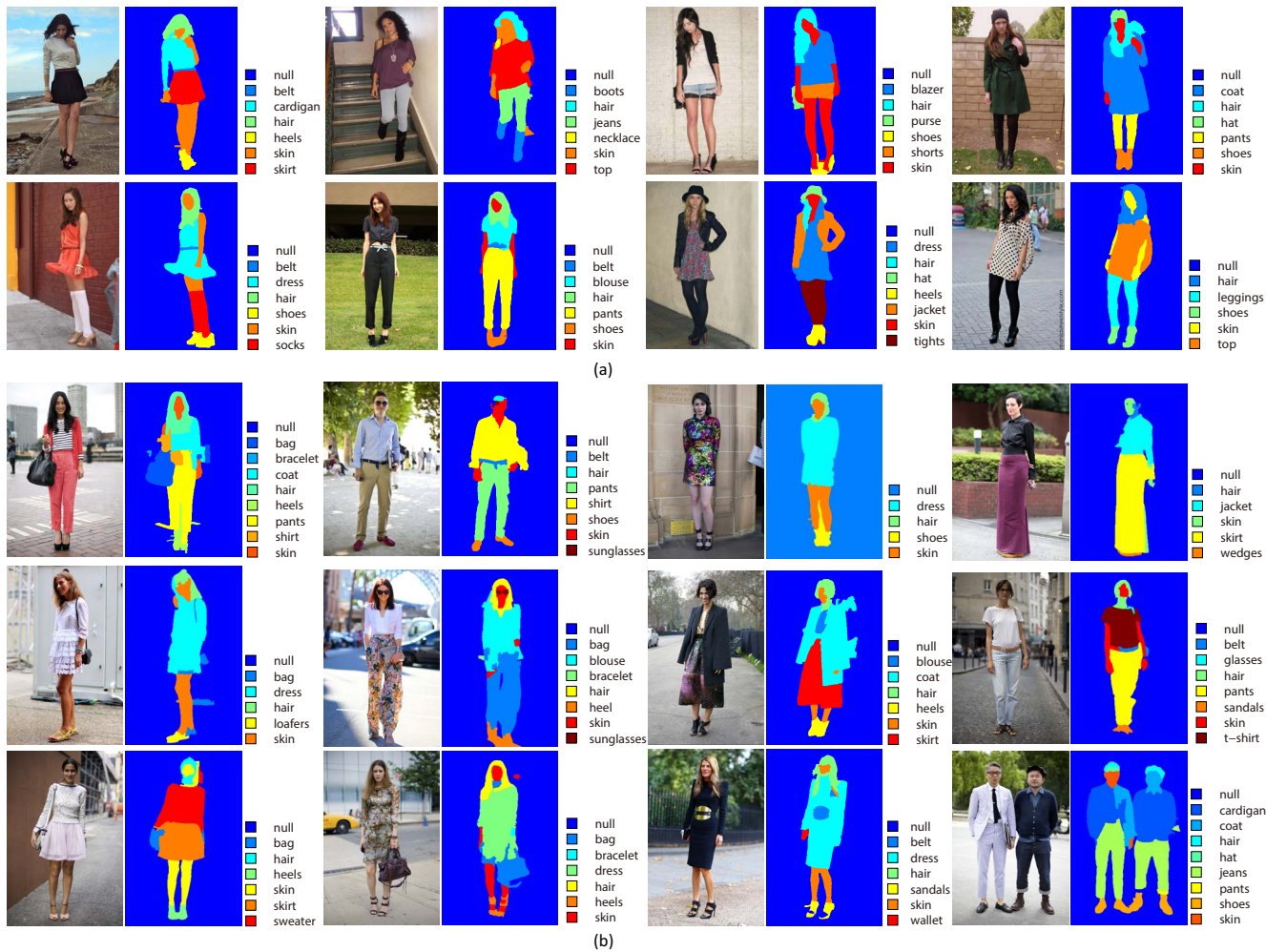
Figure 4. Some successful parsing results on (a) Fashionista (b) CCP.

tions (r1c2[3], r4c2). Moreover, our method could even 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 even parse several persons in a single image simultaneously (r5c5).

Some failure cases are shown in Figure 5. Our co-parsing framework may lead wrong results under following scenarios: (a) ambiguous patterns exist within a clogging garment item; (b) different clothing garment items share similar appearance; (c) background is extremely disordered; (d) illumination condition is poor.

4.4. Efficiency

All the experiments are carried out 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 sec per image. The co-labeling phase costs less than 1 minute 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.

5. Conclusions

This paper has proposed a well-engineered framework for joint parsing a batch of clothing images given the image-level clothing tags. Another contribution of this paper is a high-resolution street fashion photos dataset with annotations. The experiments demonstrate that our framework is effective and applicable compared with the state-of-the-art methods. In future work, we plan to improve the inference by iterating the two phases to bootstrap the results. In addition, the parallel implementation would be studied to adapt the large scale applications.
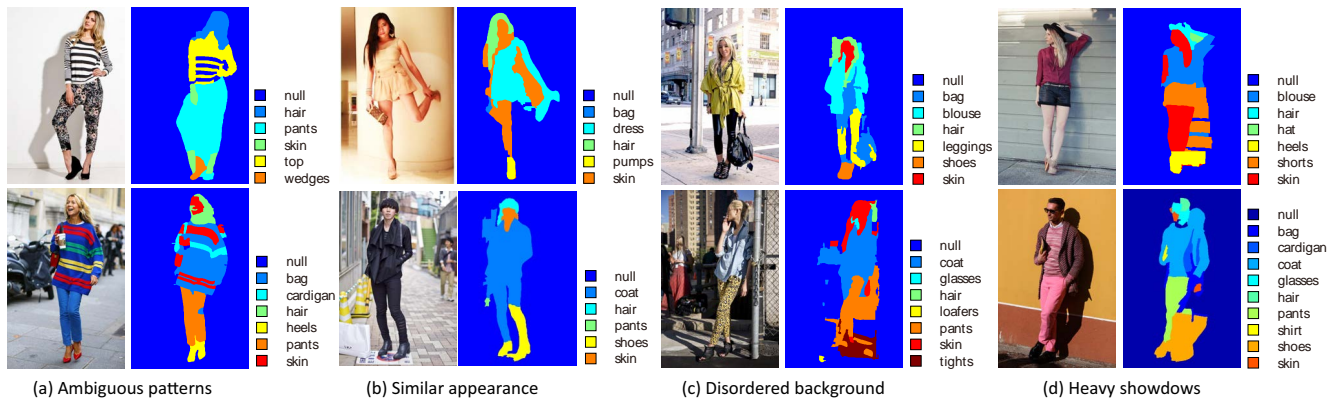
---

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

Figure 5. Some failure cases on Fashionista (1st row) and CCP (2nd row).

References

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *TPAMI*, 2011. 5

[2] Y. Bo and C. C. Fowlkes. Shape-based pedestrian parsing. In *CVPR*, 2011. 2

[3] A. Borras, F. Tous, J. Lladós, and M. Vanrell. High-level clothes description based on colour-texture and structural features. In *Pattern Recognition and Image Analysis*. Springer, 2003. 2

[4] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *TPAMI*, 2001. 2

[5] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *ECCV*, 2012. 1

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

[7] A. Gallagher and T. Chen. Clothing cosegmentation for recognizing people. In *CVPR*, 2008. 1

[8] B. Hasan and D. Hogg. Segmentation using deformable spatial priors with application to clothing. In *BMVC*, 2010. 2

[9] G. Kim, E. Xing, L. Fei-Fei, and T. Kanade. Distributed cosegmentation via submodular optimization on anisotropic diffusion. In *ICCV*, 2011. 1

[10] D. Kuettel, M. Guillaumin, and V. Ferrari. Segmentation propagation in imagenet. In *ECCV*. Springer, 2012. 2

[11] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *Workshop on Statistical Learning in Computer Vision, ECCV*, 2004. 3

[12] L. Lin, X. Liu, S. Peng, H. Chao, Y. Wang, and B. Jiang. Object categorization with sketch representation and generalized samples. *Pattern Recognition*, 2012. 1

[13] L. Lin, X. Liu, and S.-C. Zhu. Layered graph matching with composite cluster sampling. *TPAMI*, 2010. 3

[14] L. Lin, T. Wu, J. Porway, and Z. Xu. A stochastic graph grammar for compositional object representation and recognition. *Pattern Recognition*, 2009. 1

[15] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer. *TPAMI*, 2011. 3

[16] S. Liu, J. Feng, Z. Song, T. Zhang, H. Lu, C. Xu, and S. Yan. Hi, magic closet, tell me what to wear! In *ACM MM*, 2012. 1

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

[18] X. Liu, B. Cheng, S. Yan, J. Tang, T. S. Chua, and H. Jin. Label to region by bi-layer sparsity priors. In *ACM MM*, 2009. 3, 6

[19] X. Liu, L. Lin, S. Yan, H. Jin, and W. Tao. Integrating spatio-temporal context with multiview representation for object recognition in visual surveillance. *TCSVT*, 2011. 1

[20] P. Luo, X. Wang, L. Lin, and X. Tang. Joint semantic segmentation by searching for compatible-competitive references. In *ACM MM*, 2012. 3

[21] P. Luo, X. Wang, and X. Tang. Hierarchical face parsing via deep learning. In *CVPR*, 2012. 2

[22] P. Luo, X. Wang, and X. Tang. Pedestrian parsing via deep decompositional network. In *ICCV*, 2013. 2

[23] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011. 1, 2, 4

[24] S. Nowozin and S. Jegelka. Solution stability in linear programming relaxations: Graph partitioning and unsupervised learning. In *ICML*, 2009. 4

[25] K. Shi, K. Wang, J. Lu, and L. Lin. Pisa: Pixelwise image saliency by aggregating complementary appearance contrast measures with spatial priors. In *CVPR*, 2013. 4

[26] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, 2008. 6

[27] A. Vezhnevets, V. Ferrari, and J. Buhmann. Weakly supervised semantic segmentation with a multi-image model. In *ICCV*, 2011. 2

[28] N. Wang and H. Ai. Who blocks who: Simultaneous clothing segmentation for grouping images. In *ICCV*, 2011. 2

[29] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *ICCV*, 2005. 3

[30] K. Yamaguchi, H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *CVPR*, 2012. 1, 2, 6