A High Performance CRF Model for Clothes Parsing

Edgar Simo-Serra[1], Sanja Fidler[2], Francesc Moreno-Noguer[1], and Raquel Urtasun[2]

[1]IRI (CSIC-UPC) [2]University of Toronto

Abstract. In this paper we tackle the problem of clothing parsing: Our goal is to segment and classify different garments a person is wearing. We frame the problem as the one of inference in a pose-aware Conditional Random Field (CRF) which exploits appearance, figure/ground segmentation, shape and location priors for each garment as well as similarities between segments, and symmetries between different human body parts. We demonstrate the effectiveness of our approach on the Fashionista dataset [1] and show that we can obtain a significant improvement over the state-of-the-art.

1 Introduction

The impact of fashion and clothing is tremendous in our society. According to the Forbes magazine [2], excluding auctions, US online retail sales are expected to reach 262 billion dollars this year, 13% higher than the total in 2012. The situation is similar in Europe, with the expectation being that it will reach 128 billion euros. This is reflected in the growing interest in recognizing clothing from images [3–10], as this can enable a wide variety of applications such as trying on virtual garments in online shopping. Being able to automatically parse clothing is also key in order to conduct large-scale sociological studies related to family income or urban groups. For instance, several researches have attempted to estimate sociological patterns from clothing inferred from images, predicting for example occupation [11] or urban tribes [12].

In the context of fashion, Yamaguchi et al. [1], created *Fashionista*, a dataset of images and clothing segmentation labels. Great performance was obtained when the system was given information about which garment classes, but not their location, are present for each test image. Unfortunately, the performance of the state-of-the-art methods [1, 8] is rather poor when this kind of information is not provided at test time. This has been very recently partially addressed in [13] by utilizing over 300,000 weakly labeled images, where the weak annotations are in the form of image-level tags. In this paper, we show an approach which outperforms the state-of-the-art significantly without requiring these additional annotations, by exploiting the specific domain of the task: clothing a person. An example of our result can be seen in Fig. 1.

The complexity of the task of human semantic segmentation comes from the inherent variability of pose and cloth appearances, the presence of self-occlusions as well as the potentially large number of classes. Consider for example Fig. 2: an autonomous system needs to distinguish between blazers and cardigans, stockings and tights, and heels, wedges and shoes, where the intra-class variability is fundamentally much larger than the inter-class variability. This fine-grained categorization is difficult to resolve even for humans who are not familiar with the fashion industry. The problem is further aggravated by the power law distribution of classes, as certain categories have very few examples. Thus, extra-care has to be taken into account to not over-predict the classes that are very likely to appear in each image, e.g., skin, hair.

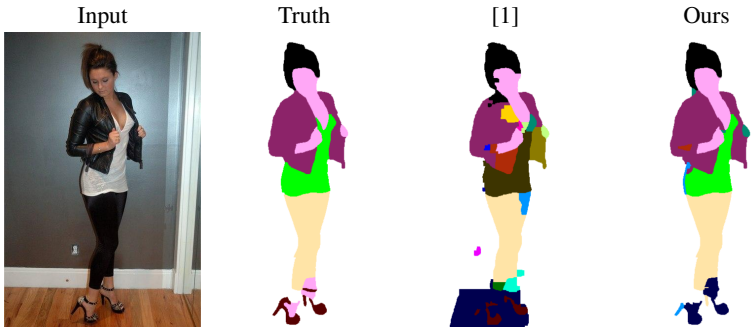


Fig. 1: Example of our result in a scenario where no a priori knowledge of which garments are worn is provided. We compare against state-of-the-art. Despite some mistakes, our result looks visually much more natural than the competing method.

Fig. 2: Examples of fine-grained annotations in Fashionista [1]. Many of the different classes are very difficult to distinguish even for humans. Observe the subtle differences between some classes such as footwear (heels, wedges, and shoes), blazer and cardigan, or stockings and tights. We also point out that this dataset has been annotated via superpixels, and thus the ground truth contains errors when superpixels do not align with the actual garments. We have not modified the ground truth segmentation in any way.

In this paper we address some of these challenges and formulate the problem as the one of inference in a Conditional Random Field (CRF), which takes into account the complex dependencies between clothing and human pose. Specifically, we develop a rich set of potentials which encode the person's global appearance and shape to perform figure/ground segmentation, shape and location likelihoods for each garment, which we call *clothelets*, and long-range similarity between segments to encourage, for example, T-Shirt pixels on the body to agree with the T-shirt pixels on the person's arm. We further exploit the fact the people are symmetric and dress as such as well by introducing symmetry-based potentials between different limbs. We also use a variety of different local features encoding cloth appearance as well as local shape of the person's parts. We demonstrate the effectiveness of our approach of the Fashionista dataset [1] and show that our approach significantly outperforms the existing state-of-the-art.

2 Related Work

There has been a growing interest in recognizing outfits and clothing from still images. One of the first approaches on the subject was Chen et al. [14], which manually built a composite clothing model, that was then matched to input images. This has led to more recent applications for learning semantic clothing attributes [5], which are in turn used for describing and recognizing the identity of individuals [4, 6], their style [3], and performing sociological studies such as predicting the occupation [11] or urban tribes [12]. Other tasks like outfit recommendations [15] have also been investigated. However, in general, these approaches do not perform accurate segmentation of clothing, which is the goal of our approach. Instead, they rely on more coarse features such as bounding boxes and focus on producing generic outputs based on the presence/absence of a specific type of outfit. It is likely that the performance of such systems would improve if accurate clothing segmentation would be possible.

Recent advances in 2D pose estimation [16, 17] have enabled a more advanced segmentation of humans [18]. However, most approaches have focused on figure/ground labeling [19, 20]. Additionally, pose information has been used as a feature in clothing related tasks such as finding similar worn outfits in the context of online shopping [9].

Segmentation and classification of garments has been addressed in the restrictive case in which the labels are known beforehand [1]. The original paper tackled this problem in the context of fashion photographs which depicted one person typically in an upright pose. This scenario also been extended to the case where more than one individual can be present in the image [8]. In order to perform the segmentation, conditional random fields are used with potentials linking clothing and pose. However, the performance of these approaches drops significantly when no information about the outfit is known a priori (i.e., no tags are provided at test time). The paper doll approach [13] uses over 300,000 weakly labeled training images and a small set of fully labeled examples in order to enrich the model of [1] with a prior over image labels. As we will show in the experimental evaluation, our method can handle this scenario without having to resort to additional training images. Furthermore, it consistently outperforms [1, 13].

CRFs have been very successful in semantic segmentation tasks. Most approaches combine detection and segmentation by using detectors as additional image evidence [21, 22]. Co-occurrence potentials have been employed to enforce consistency among region labels [23]. Part-based detectors have also been aligned to image contours to aid in object segmentation [24]. All these strategies have been applied to generic segmentation problems, where one is interested in segmenting classes such as car, sky or trees. Pixel-wise labeling of clothing is, however, a much more concrete task, where strong domain specific information, such as 2D body pose, can be used to reduce ambiguities.

3 Clothing a Person

We pose the clothing parsing problem as one of inference in a Conditional Random Field (CRF), which takes into account complex dependencies that exist between garments and human pose. We obtain pose by employing a 2D articulated model by Yang et al. [17] which predicts the main keypoints such as head, shoulders, knees, etc. As [1], we will exploit these keypoints to bias the clothing labeling in a plausible way (e.g., a

Table 1: Overview of the different types of potentials used in the proposed CRF model.

Type	Name	Description
unary	Simple features ($\phi_{i,j}^{simple}(y_i)$)	Assortment of simple features [1].
unary	Object mask ($\phi_{i,j}^{obj}(y_i)$)	Figure/ground segmentation ask.
unary	Clothelets ($\phi_{i,j}^{cloth}(y_i)$)	Pose-conditioned garment likelihood masks.
unary	Ranking ($\phi_{i,j}^{o2p}(y_i)$)	Rich set of region ranking features.
unary	Bias ($\phi_j^{bias}(y_i)$ and $\phi_{p,j}^{bias}(l_p)$)	Class biases.
pairwise	Similarity ($\phi_{m,n}^{simil}(y_m, y_n)$)	Similarity between superpixels.
pairwise	Compatibility ($\phi_{i,p}^{comp}(y_i, l_p)$)	Edges between limb segments and superpixels.

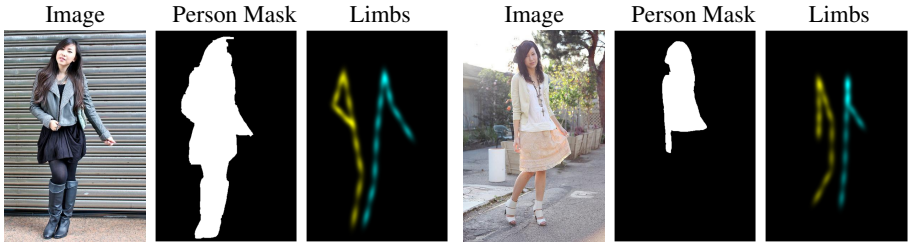


Fig. 3: Visualization of CPMC object segments [26] and limbs (obtained via [17]). Note that CPMC typically generates high quality results, e.g. the one in the left image, but can also completely miss large parts of the body as shown in the image on the right.

hat is typically on the head and not the feet). To manage the complexity of the segmentation problem we represent each input image with a small number of superpixels [25]. Our CRF contains a variable encoding the garment class (including background) for each superpixel. We also add limb variables which encode the garment associated with a limb in the human body and correspond to edges in the 2D articulated model. We use the limb variables to propagate information while being computationally efficient.

Our CRF contains a rich set of potentials which exploit the domain of the task. We use the person's global appearance and shape to perform figure/ground segmentation in order to narrow down the scope of cloth labeling. We further use shape and location likelihoods for each garment, which we call *clothelets*. We exploit the fact that people are symmetric and typically dress as such by forming long-range consistency potentials between detected symmetric keypoints of the human pose. We finally also use a variety of different features that encode appearance as well as local shape of superpixels.

3.1 Pose-aware Model

Given an input image represented with superpixels, our goal is to assign a clothing label (or background) to each of them. More formally, let $y_i \in \{1, \cdots, C\}$ be the class associated with the i-th superpixel, and let l_p be the p-th limb segment defined by the edges in the articulated body model. Each limb l_p is assumed to belong to one class, $l_p \in \{1, \cdots, C\}$. To encode body symmetries in an efficient manner, we share limb variables between the left and right part of the human body, e.g., the left and the right leg share the same limb variables. We propose several domain inspired potentials, the overview of which is presented in Table 1. We emphasize that the weights associated with each potential in our CRF will be learned using structure prediction. We now explain each potential in more detail.

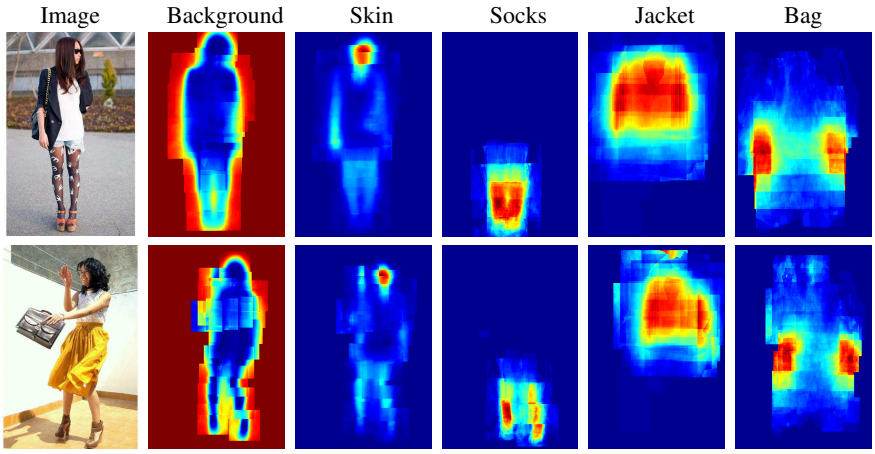


Fig. 4: Visualization of different clothelets for two different input images.

Simple Features: Following [1], we concatenate a diverse set of simple local features and train a logistic regression classifier for each class. In particular, we use color features, normalized histograms of RGB and CIE L*a*b* color; texture features, Gabor filter responses; and location features: both normalized 2D image coordinates and pose-relative coordinates. The output of the logistic functions are then used as unary features in the CRF. This results in a unary potential with as many dimensions as classes:

$$\phi_{i,j}^{simple}(y_i) = \begin{cases} \sigma_j^{simple}(f_i), & \text{if } y_i = j \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where $\sigma_j^{simple}(f_i)$ is the score of the classifier for class j, and f_i is the concatenation of all the features for superpixel i. Notice we have used C different unary potentials, one for each class. By doing this, we allow the weights of a variety of potentials and classes be jointly learned within the model.

Figure/Ground Segmentation: To facilitate clothing parsing we additionally compute how likely each superpixel belongs to a person. We do this by computing a set of bottom-up region proposals using the CPMC approach [26]. We take top K (we set $K = 100$) regions per image and use O2P [27] to score each region into figure/ground (person-vs-background). Since we know that there is a person in each image, we take at least the top scoring segment per image, no matter its score. For images with multiple high scoring segments, we take the union of all segments with scores higher than a learned threshold [27]. We define a unary potential to encourage the superpixels that lie inside the foreground mask to take any of the clothing labels (and not background):

$$\phi_{i,j}^{obj}(y_i) = \begin{cases} \sigma^{cpmc} \cdot |\neg M_{fg} \cap S_i|/|S_i|, & \text{if } y_i = 1 \\ \sigma^{cpmc} \cdot |M_{fg} \cap S_i|/|S_i|, & \text{otherwise} \end{cases} \quad (2)$$

where $y_i = 1$ encodes the background class, σ^{cpmc} is the score of the foreground region, S_i, M_{fg} are binary masks defining the superpixel and foreground, respectively, and $\neg M_{fg}$ is a mask of all pixels not in foreground. Fig. 3 shows examples of masks

obtained by [27]. Note that while in some cases it produces very accurate results, in others, it performs poorly. These inaccurate masks are compensated by other potentials.

Clothelets: Our next potential exploits the statistical dependency between the location on the human body and garment type. Its goal is to make use of the fact that e.g. jeans typically cover the legs and not the head. We compute a likelihood of each garment appearing in a particular relative location of the human pose. In particular, for each training example we take a region around the location of each joint (and limb), the size of which corresponds to the size of the joint part template encoded in [17]. We average the GT segmentation masks for each class across the training examples. In order to capture garment classes that stray away from the pose, we use boxes that are larger than the part templates in [17]. At test time, the masks for each class are overlaid relative to the inferred pose and normalized by the number of non-zero elements. Areas with no information are assigned to the background class. The potential is then defined as

$$\phi_{i,j}^{cloth}(y_i) = \begin{cases} (\text{clothelet}_i^j \cdot S_i)/|S_i|, & \text{if } y_i = j \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where clothelet_i^j is the clothelet for the j-th class, and $\cdot$ is the dot product. Fig. 4 depicts clothelets for a few sample classes.

Shape Features: This potential uses a set of rich features that exploit both the shape and local appearance of garments. In particular, we use eSIFT and eMSIFT proposed by [27] for region description. Given a region, both descriptors extract SIFT inside the region and enrich it with the relative location and scale within the region. Second-order pooling is used to define the final region descriptor. eSIFT and eMSIFT differ slightly in how the descriptors are pooled, eSIFT pools over both the region and background of the region, while eMSIFT pools over the region alone. While [27] defines the features over full object proposals, here we compute them over each superpixel. As such, they capture more local shape of the part/limb and local appearance of the garment. We train a logistic classifier for each type of feature and class and use the output as our potential:

$$\phi_{i,j}^{o2p}(y_i) = \begin{cases} \sigma_j^{o2p}(r_i), & \text{if } y_i = j \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

with $\sigma_j^{o2p}(r_i)$ the classifier score for class j, and r_i the feature vector for superpixel i.

Bias: We use a simple bias for the background to encode the fact that it is the class that appears more frequently. Learning a weight for this bias is equivalent to learning a threshold for the foreground, however within the full model. Thus:

$$\phi^{bias}(y_i) = \begin{cases} 1, & \text{if } y_i = 1 \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

Similarity: In CRFs, neighboring superpixels are typically connected via a (contrast-sensitive) Potts model encouraging smoothness of the labels. For clothing parsing, we want these connections to act on a longer range. That is, a jacket is typically split in

multiple disconnected segments due to a T-shirt, tie, and/or a bag. Our goal is to encourage superpixels that are similar in appearance to agree on the label, even though they may not be neighbors in the image.

We follow [28] and use size similarity, fit similarity that measures how well two superpixels fit each other; and color and texture similarity, with the total of 12 similarity features between each pair of superpixels. We then train a logistic regression to predict if two superpixels should have the same label or not. In order to avoid setting connections on the background, we only connect superpixels that overlap with the bounding box of the 2D pose detection. Note that connecting all pairs of similar superpixels would slow down inference considerably. To alleviate this problem, we compute the minimum spanning tree using the similarity matrix and use the top 10 edges to connect 10 pairs of superpixels in each image. We form a pairwise potential between each connected pair:

$$\phi^{simil}_{m,n}(y_m, y_n) = \begin{cases} \sigma^{simil}_{m,n}, & \text{if } y_m = y_n \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where $\sigma^{simil}_{m,n}$ is the output of the similarity classifier.

Limb Segment Bias: We use a per-class bias on each limb segment to capture a location specific bias, e.g., hat only appears in the head:

$$\phi^{bias}_{p,j}(l_p) = \begin{cases} 1, & \text{if } l_p = j \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

These potentials allow us to compute which classes are more frequent in each limb.

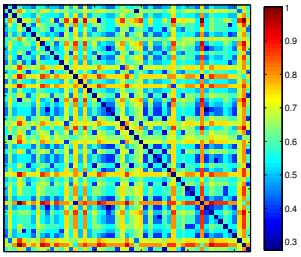
Compatibility Segmentation-Limbs: We define potentials connecting limb segments with nearby superpixels encouraging them to agree in their labels. Towards this goal, we first define a Gaussian mask centered between two joints. More formally, for two consecutive joints with coordinates $J_a = (u_a, v_a)$ and $J_b = (u_b, v_b)$, we define the mask based on the following Normal distribution:

$$M(J_a, J_b) = \mathcal{N}\left(\frac{J_a + J_b}{2}, \quad R\begin{pmatrix} q_1\|J_a - J_b\| & 0 \\ 0 & q_2 \end{pmatrix} R^{\mathrm{T}}\right) \tag{8}$$

where R is a 2D rotation matrix with an angle $\arctan(\frac{u_a - u_b}{v_a - v_b})$, and q_1 and q_2 are two hyperparameters controlling the spread of the mask longitudinally and transversely, respectively. The strength of the connection is based on the overlap between the superpixels and the Gaussian mask:

$$\phi^{comp}_{i,p}(y_i, l_p) = \begin{cases} M(J_a, J_b) \cdot S_i, & \text{if } y_i \neq 1 \text{ and } y_i = k_p \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

For computational efficiency, edges with connection strengths below a threshold are not set in the model. Some examples of the limb segment masks are shown in Fig. 3. We can see the masks fit the body tightly to avoid overlapping with background superpixels.



Clothing classes			Similarity (d_{lch})
jumper	vs	sweater	1.000
jumper	vs	sweatshirt	0.697
jumper	vs	top	0.557
jumper	vs	stockings	0.507
jumper	vs	heels	0.465
jumper	vs	sunglasses	0.341
jumper	vs	watch	0.317

Fig. 5: Inverse of the Leacock-Chodorow Similarity between the classes in the Fashionista dataset. We display the similarity matrix between all the classes on the left. Some individual values of the similarity between the jumper class and several other classes can be seen on the right.

Full Model: We define the energy of the full model to be the sum of three types of energies encoding unary and pairwise potentials that depend on the superpixel labeling, as well as an energy term linking the limb segments and the superpixels:

$$E(\mathbf{y},\mathbf{l}) = E_{unary}(\mathbf{y}) + E_{similarity}(\mathbf{y}) + E_{limbs}(\mathbf{y},\mathbf{l}) \tag{10}$$

This energy is maximized during inference. The unary terms are formed by the concatenation of appearance features, figure/ground segmentation, clothelets, shape features and background bias for a total of $K = (1 + 5C)$ features

$$E_{unary}(\mathbf{y}) = \sum_{i=1}^{N}\sum_{j=1}^{K} \mathbf{w}_j^{unaries}\phi_{i,j}^{unary}(y_i) \tag{11}$$

where N is the number of superpixels. The pairwise features encode the similarity between different pairs of superpixels as we describe above

$$E_{similarity}(\mathbf{y}) = \sum_{(m,n)\in\text{pairs}} w^{simil}\phi_{m,n}^{simil}(y_m, y_n) \tag{12}$$

The limb-superpixel compatibility term is defined as

$$E_{limbs}(\mathbf{y},\mathbf{l}) = \sum_{p=1}^{M}\left(\sum_{j=1}^{C}\left(w_j^{bias}\phi_{p,j}^{bias}(l_p) + \sum_{i=1}^{N} w_{j,p}^{comp}\phi^{comp}(y_i, l_p)\right)\right) \tag{13}$$

for a total of $(M + C)$ features, with M the number of limb segments.

3.2 Learning and Inference

Our model is a multi-label CRF which contains cycles and thus inference is NP-hard. We use a message passing algorithm, distributed convex belief propagation [29] to perform inference. It belongs to the set of LP-relaxation approaches, and has convergence guarantees. This is not the case in other message passing algorithms such as loopy-BP.

To learn the weights, we use the primal-dual method of [30] (we use the implementation of [31]), shown to be more efficient than other structure prediction learning algorithms. As loss-function, we use the semantic similarity between the different classes

Table 2: Comparison against the state-of-the-art on two different datasets: Fashionista v0.2 with 56 classes and Fashionista v0.3 with 29 classes.

	29 Classes		56 Classes		
Method	[1]	Ours	[1]	[13]	Ours
Jaccard index	12.32	**20.52**	7.22	9.22	**12.28**

in order to penalize mistakes between unrelated classes more than similar ones. We do this via Wordnet [32], which is a large lexical database in which sets of cognitive synonyms (synsets) are interlinked by means of semantic and lexical relationships. We can unambiguously identify each of the classes with a single synset, and then proceed to calculate similarity scores between these synsets that represent the semantic similarity between the classes, in order to penalize mistakes with dissimilar classes more.

In particular, we choose the corpus-independent Leacock-Chodorow Similarity score. This score takes into account the shortest path length p between both synsets and the maximum depth of the taxonomy d at which they occur. It is defined as the relationship $-\log(p/2d)$. A visualization of the dissimilarity between all the classes in the dataset can be seen in Fig. 5. We therefore define the loss-function as:

$$\Delta^y(y_i, y_i^*) = \begin{cases} 0, & \text{if } y_i = y_i^* \\ d_{\text{lch}}(y_i, y_i^*), & \text{otherwise} \end{cases} \tag{14}$$

with $d_{\text{lch}}(\cdot,\cdot)$ being the inverse Leacock-Chodorow Similarity score between both classes. For the limb segments we use a 0-1 loss:

$$\Delta^k(k_i, k_i^*) = \begin{cases} 0, & \text{if } k_i = k_i^* \\ 1, & \text{otherwise} \end{cases} \tag{15}$$

4 Experimental Evaluation

We evaluate our approach on both a the Fashionista dataset v0.3 [1], and the setting of [13] with the Fashionista dataset v0.2. Both datasets are taken from http://www.chictopia.com in which a single person appears wearing a diverse set of garments. The dataset provides both annotated superpixels as well as 2D pose annotations. A set of evaluation metrics and the full source code of approaches [1, 13] are provided. Version 0.2 has 685 images and v0.3 has 700 images. Note that v0.3 is not a superset of v0.2.

We have modified the Fashionista v0.3 dataset in two ways. First we have compressed the original 54 classes into 29. This is due to the fact that many classes that appear have very few occurrences. In fact, in the original dataset, 13 classes have 10 or fewer examples and 6 classes have 3 or fewer instances. This means that when performing a random split of the samples into training and test subsets, there is a high probability that some classes will only appear in one of the subsets. We therefore compress the classes by considering both semantic similarity and the number of instances. The final classes in this setting can be seen in the supplemental material of this paper.

For evaluation on Fashionista v0.3 we consider a random 50-50 train-test split. As previously stated, we do not have information about which classes are present in the scene. We employ the publicly available code of [1] as the baseline. We evaluate on

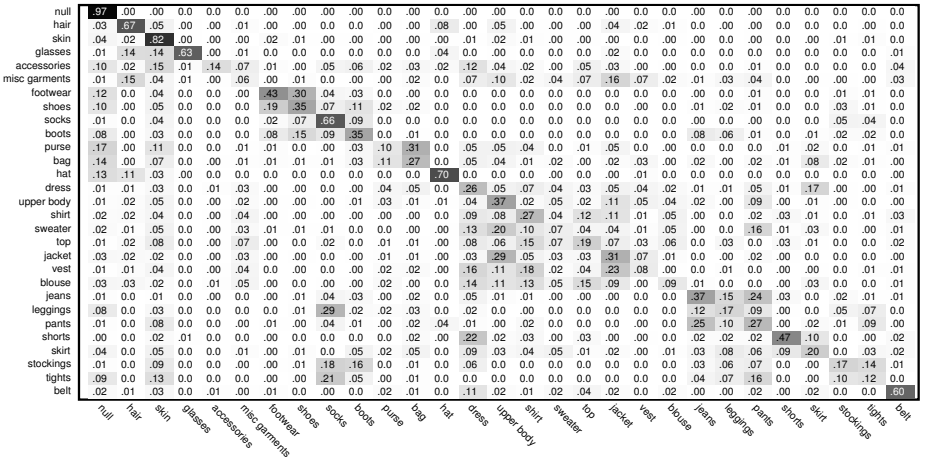


Fig. 6: Confusion matrix for our approach on the Fashionista v0.3 dataset.

Table 3: Evaluation on foreground segmentation task on the Fashionista v0.2 dataset.

Method	CPMC [27]	[1]	Clothelets	[13]	Ours
Pixel Accuracy	-	77.98	77.09	84.68	**84.88**
Person/Bck. Accuracy	85.39	93.79	94.77	95.79	**97.37**

Fashionista v0.2 according to the methodology in [13]. This consists of a split with 456 images for training and 229 images for testing. Note that [13] uses 339,797 additional weakly labeled images from the Paper doll dataset for training, which we do not use.

Following PASCAL VOC, we report the average class intersection over union (Jaccard index). This metric is the most similar to human perception as it considers all true positives, true negatives and false positives. It is nowadays a standard measure to evaluate segmentation and detection [33–35].

Comparison to State-of-the-Art: We compare our approach against [1, 13]. The approach of [1] uses a CRF with very simple features. We adapt the code to run in the setting in which the labels that appear in the image are not known a priori. Note also that [13] uses a look-up approach on a separate dataset to parse the query images. The results of the comparison can be seen in Table 2. Note that our approach consistently outperforms both competing methods on both datasets, even though [13] uses 339,797 additional images for training. We roughly obtain a 60% relative improvement on Jaccard index metric with respect to [1] and a 30% improvement over [13]. The full confusion matrix of our method can be seen in Fig. 6. We can identify several classes that have large appearance variation and similar positions that get easily confused, such as Footwear with Shoes and Jeans with Pants.

Foreground Segmentation: We also evaluate person-background segmentation results. Note that the binary segmentation in our model is obtained by putting all foreground garment classes to the person class. In Table 3, we show results for both pixel accuracy considering all the different classes, and the two class case of foreground/background segmentation accuracy. We see that the best results are obtained by the approaches rea-

Table 4: Influence of pose. We compare against the state-of-the-art in three different scenarios: estimated 2D pose, ground truth 2D pose and no pose information at all.

Method		29 Classes	56 Classes
[1]	Estimated	12.32	7.22
	GT Pose	12.39	7.41
	No Pose	10.54	5.22
Ours	Estimated	20.52	12.28
	GT Pose	21.01	12.46
	No Pose	16.56	9.64

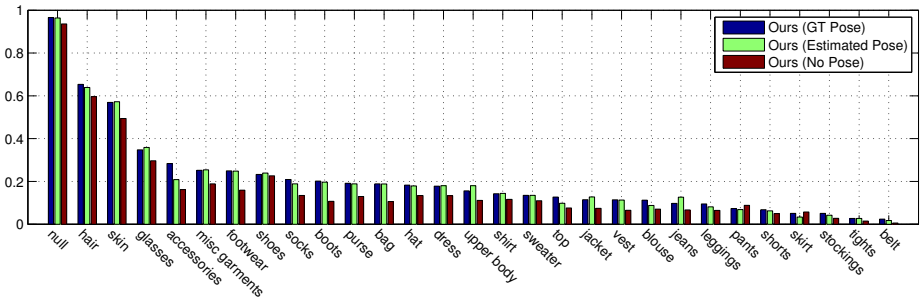


Fig. 7: Per class results for our model using Jaccard index metric on Fashionista v0.3.

Table 5: Oracle performance for different of superpixels for the Fashionista v0.3 dataset.

Threshold	0.16	0.10	0.05
Mean superpixels/image	50	120	290
Jaccard index	69.44	83.07	100

soning jointly about the person and clothing. Our approach outperforms the baseline CPMC [27] by 12%, and achieves a 4% over [1] and 2% over [13].

Pose Influence: We next investigate the importance of having an accurate pose estimate. Towards this goal, we analyze three different scenarios. In the first one, the pose is estimated by [17]. The second case uses the ground-truth pose, while the last one does not use pose information at all. As shown in Table 4, the poses in this dataset are not very complex as performance does not increase greatly when using ground truth instead of estimated pose. However, without pose information, performance drops 20%. This shows that our model is truly pose-aware. A breakdown of the effect of pose on all the classes is shown in Fig. 7. Some classes like hat, belt or boots benefit greatly from pose information while others like shorts, tights or skin do not really change.

Oracle Performance: Unlike [1], we do not use the fine level superpixels, but instead use coarser superpixels to speed up learning and inference. Table 5 shows that using coarser superpixels lowers the maximum achievable performance. However, by having larger areas, the local features become more discriminative. We also note that the dataset [1] was annotated by labeling the finer superpixels. As some superpixels do not follow boundaries well, the ground truth contains a large number of errors. We did not correct those, and stuck with the original annotations.

Table 6: Different results using only unary potentials in our model.

Method	Simple features [1]	Clothelets	eSIFT [27]	eMSIFT [27]
29 Classes	13.80	8.91	16.65	13.65
56 Classes	7.93	3.02	9.29	7.80

Table 7: Importance of the different potentials in our model in the 56 class setting.

Method	Jaccard index
Full Model	12.28
No similarity ($\phi_{m,n}^{simil}(y_m, y_n)$)	11.64
No limb segments ($\phi_{i,p}^{comp}(y_i, l_p)$)	12.24
No simple features ($\phi_{i,j}^{simple}(y_i)$)	10.07
No clothelets ($\phi_{i,j}^{cloth}(y_i)$)	11.94
No object mask ($\phi_{i,j}^{obj}(y_i)$)	10.02
No eSIFT ($\phi_{i,j}^{o2p(eSIFT)}(y_i)$)	10.70
No eMSIFT ($\phi_{i,j}^{o2p(eMSIFT)}(y_i)$)	12.25

Importance of the Features: We also evaluate the influence of every potential in our model in Table 6. The eSIFT features obtain the best results under the Jaccard index metric. The high performance of eSIFT can be explained by the fact that it also takes into account the super pixel's background, thus capturing local context of garments. This feature alone surpasses the simple features from [1] despite that it does not use pose information. By combining all the features we are able to improve the results greatly. We show some qualitative examples of the different feature activations in Fig. 8. We also evaluate the model in a leave-one-out fashion. That is, for each unary we evaluate the rest of the unaries in the model without it. Results are shown in Table 7.

Qualitative Results: We show qualitative results for both our approach and the current state-of-the-art in Fig. 9. We can see a visible improvement over [1], especially on person/bckgr classification due to the strength of the proposed clothelets and person masks which in combination give strong cues on person segmentation. There are also several failure cases of our algorithm. One of the main failure cases is a breakdown of the superpixels caused by clothing texture. An excess of texture leads to an oversegmentation where the individual superpixels are no longer discriminative enough to individually identify (Fig. 9-bottom-left), while too much similarity with the background leads to very large superpixels that mix foreground and background. Additionally it can be seen that failures in pose detection can lead to missed limbs (Fig. 9-bottom-right).

Computation Time: Our full model takes several hours to train and roughly 20 minutes to evaluate on the full test set (excluding feature computation), on a single machine. On the same machine [1] takes more than twice the time for inference. Additionally, [1] uses grid-search for training, which does not scale to a large amount of weights, that as we have shown, are able to provide an increase in performance. Even with only 2 weights, [1] is several times slower to train than our model. Furthermore, [13] reports a training time of several days in a distributed environment.

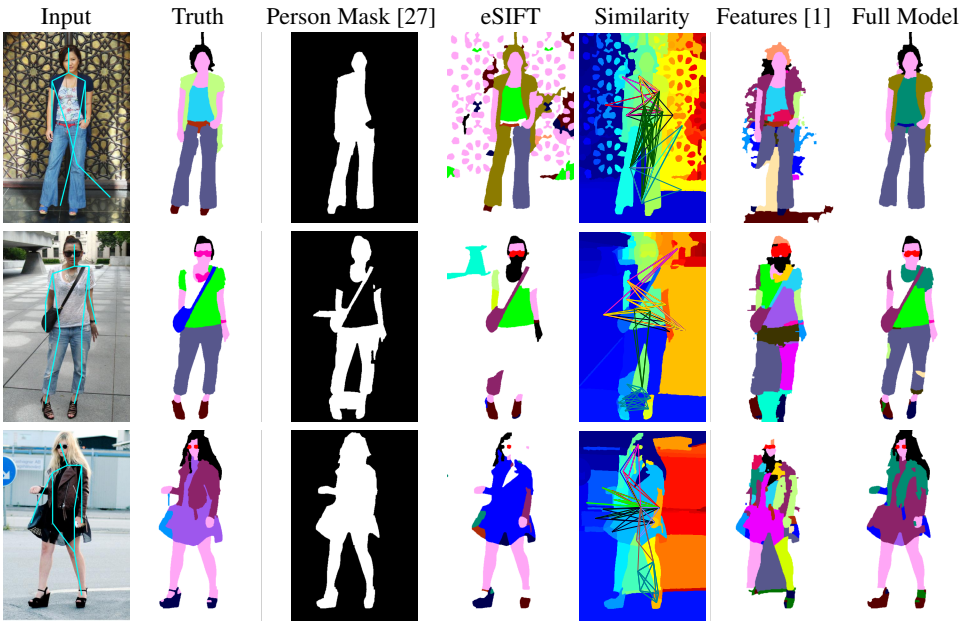


Fig. 8: We show various feature activations for example images. For similarity we display the connections between the superpixels. While both the features from [1] and eSIFT provide decent segmentation results, they have poorly defined boundaries. These are corrected via person masks [27] and clothelets. Further corrections are obtained by pairwise potentials such as symmetry and similarity. These results highlight the importance of combining complementary features. For class colors we refer to Fig. 9.

5 Conclusions

We have tackled the challenging problem of clothing parsing. We have shown that our approach is able to obtain a large improvement over the state-of-the-art in the challenging Fashionista dataset by exploiting appearance, figure/ground segmentation, shape and location priors for each garment as well as similarity between segments and symmetries between different human body parts. Despite these promising results, we believe much can still be done to improve. For example, one of the most occurring mistakes are missing glasses or other small garments. We believe a multi-resolution approach is needed to handle the diversity of garment classes. Additionally, we believe that using 3D pose instead of 2D, e.g [36], would be beneficial to handle self-occlusions better.

Acknowledgements. This work has been partially funded by Spanish Ministry of Economy and Competitiveness under projects PAU+ DPI2011-27510 and ERA-Net Chistera project ViSen PCIN-2013-047.

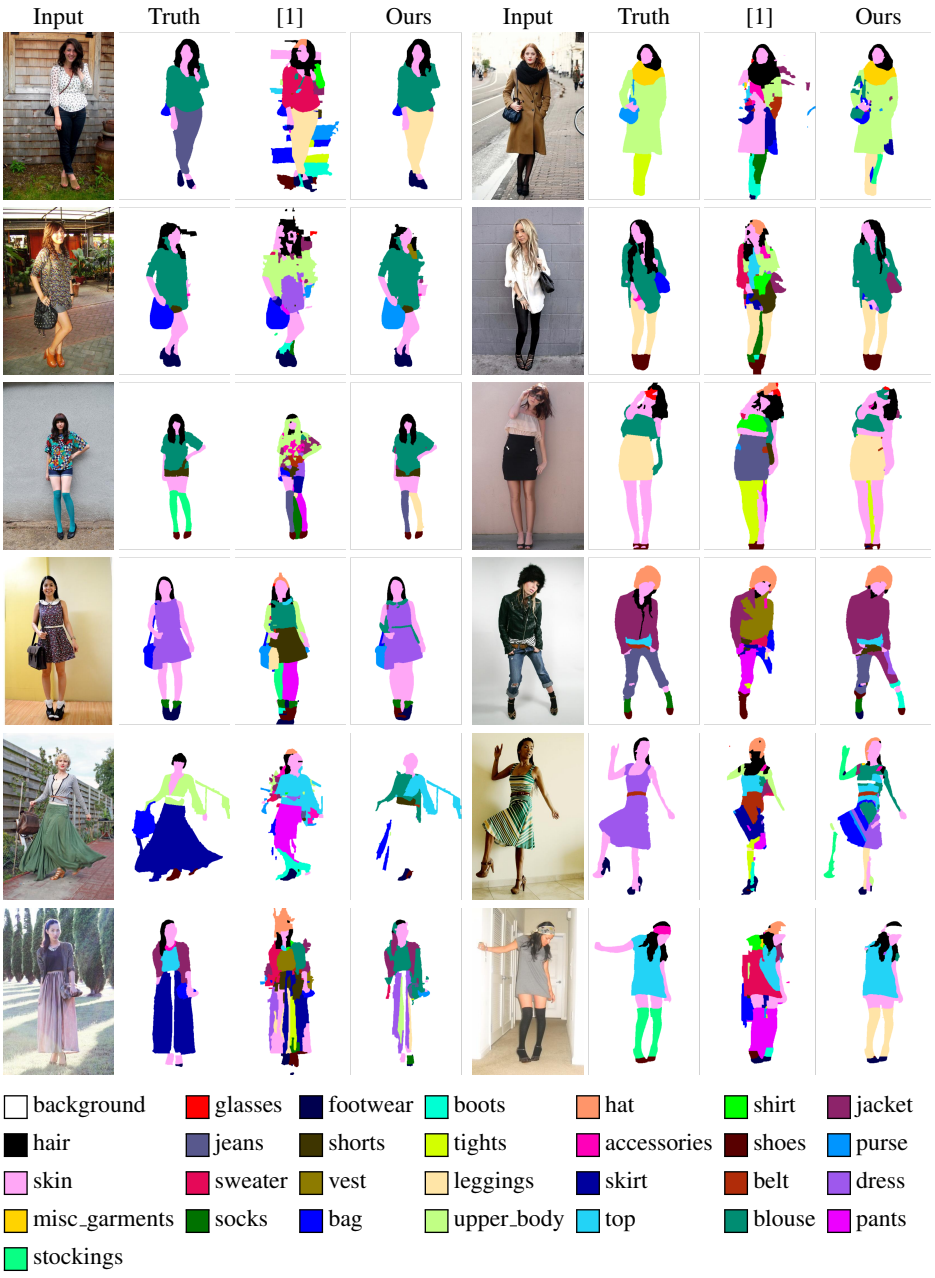


Fig. 9: Results on Fashionista v0.2 with 29 classes, comparing our approach with the state-of-the-art. In the top four rows we show good results obtained by our model. In the bottom two rows we show failure cases that are in general caused by 2D pose estimation failure, superpixel failures or chain failures of too many potentials.

References

1. Yamaguchi, K., Kiapour, M.H., Ortiz, L.E., Berg, T.L.: Parsing clothing in fashion photographs. In: CVPR. (2012)
2. Forbes Magazine: US Online Retail Sales To Reach $370B By 2017; €191B in Europe. http://www.forbes.com (2013) [Online; accessed 14-March-2013].
3. Bossard, L., Dantone, M., Leistner, C., Wengert, C., Quack, T., Gool, L.V.: Apparel classifcation with style. In: ACCV. (2012)
4. Bourdev, L., Maji, S., Malik, J.: Describing people: A poselet-based approach to attribute classification. In: ICCV. (2011)
5. Chen, H., Gallagher, A., Girod, B.: Describing clothing by semantic attributes. In: ECCV. (2012)
6. Gallagher, A.C., Chen, T.: Clothing cosegmentation for recognizing people. In: CVPR. (2008)
7. Hasan, B., Hogg, D.: Segmentation using deformable spatial priors with application to clothing. In: BMVC. (2010)
8. Jammalamadaka, N., Minocha, A., Singh, D., Jawahar, C.: Parsing clothes in unrestricted images. In: BMVC. (2013)
9. Liu, S., Song, Z., Liu, G., Xu, C., Lu, H., Yan, S.: Street-toshop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In: CVPR. (2012)
10. Wang, N., Ai, H.: Who blocks who: Simultaneous clothing segmentation for grouping images. In: ICCV. (2011)
11. Song, Z., Wang, M., s. Hua, X., Yan, S.: Predicting occupation via human clothing and contexts. In: ICCV. (2011)
12. Murillo, A.C., Kwak, I.S., Bourdev, L., Kriegman, D., Belongie, S.: Urban tribes: Analyzing group photos from a social perspective. In: CVPR Workshops. (2012)
13. Yamaguchi, K., Kiapour, M.H., Berg, T.L.: Paper doll parsing: Retrieving similar styles to parse clothing items. In: ICCV. (2013)
14. Chen, H., Xu, Z.J., Liu, Z.Q., Zhu, S.C.: Composite templates for cloth modeling and sketching. In: CVPR. (2006)
15. Liu, S., Feng, J., Song, Z., Zhang, T., Lu, H., Changsheng, X., Yan, S.: Hi, magic closet, tell me what to wear! In: Proceedings of the 20th ACM international conference on Multimedia. (2012)
16. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. In: ICCV. (2009)
17. Yang, Y., Ramanan, D.: Articulated pose estimation using flexible mixtures of parts. In: CVPR. (2011)
18. Dong, J., Chen, Q., Xia, W., Huang, Z., Yan, S.: A deformable mixture parsing model with parselets. In: ICCV. (2013)
19. Ladicky, L., Torr, P.H.S., Zisserman, A.: Human pose estimation using a joint pixel-wise and part-wise formulation. In: CVPR. (2013)
20. Wang, H., Koller, D.: Multi-level inference by relaxed dual decomposition for human pose segmentation. In: CVPR. (2011)
21. Yao, Y., Fidler, S., Urtasun, R.: Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In: CVPR. (2012)
22. Fidler, S., Sharma, A., Urtasun, R.: A sentence is worth a thousand pixels. In: CVPR. (2013)
23. Ladicky, L., Russell, C., Kohli, P., Torr, P.H.S.: Graph cut based inference with co-occurrence statistics. In: ECCV. (2010)
24. Brox, T., Bourdev, L., Maji, S., Malik, J.: Object segmentation by alignment of poselet activations to image contours. In: CVPR. (2011)

25. P.Arbelaez, M.Maire, C.Fowlkes, J.Malik: Contour detection and hierarchical image segmentation. In: PAMI. (2011)
26. Carreira, J., Sminchisescu, C.: CPMC: Automatic Object Segmentation Using Constrained Parametric Min-Cuts. TPAMI **34** (2012) 1312–1328
27. Carreira, J., Caseiroa, R., Batista, J., Sminchisescu, C.: Semantic segmentation with second-order pooling. In: ECCV. (2012)
28. Uijlings, J.R.R., van de Sande, K.E.A., Gevers, T., Smeulders, A.W.M.: Selective search for object recognition. IJCV **104** (2013) 154–171
29. Schwing, A., Hazan, T., Pollefeys, M., Urtasun, R.: Distributed message passing for large scale graphical models. In: CVPR. (2011)
30. Hazan, T., Urtasun, R.: A primal-dual message-passing algorithm for approximated large scale structured prediction. In: NIPS. (2010)
31. Schwing, A.G., Hazan, T., Pollefeys, M., Urtasun, R.: Efficient structured prediction with latent variables for general graphical models. In: ICML. (2012)
32. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38** (1995) 39–41
33. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision **88** (2010) 303–338
34. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. International Journal of Robotics Research (IJRR) (2013)
35. Deng, J., Dong, W., Socher, R., jia Li, L., Li, K., Fei-fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. (2009)
36. Simo-Serra, E., Quattoni, A., Torras, C., Moreno-Noguer, F.: A Joint Model for 2D and 3D Pose Estimation from a Single Image. In: CVPR. (2013)